Technical Progress Report on

Development of Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages

Submitted to

DEPARTMENT OF INFORMATION TECHNOLOGY
Ministry of Communication and Information Technology
Govt. of India, New Delhi

By
A Consortium Consisting of

- IIIT Hyderabad (Overall Coordination)
- IIT Kanpur
- Thapar University Patiala
- IIT Guwahati
- Tezpur University
- North Eastern Hill University (NEHU) Shillong
- Rajiv Gandhi Institute of Technology (RIT) Kottayam
- Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT) Gandhinagar
- IIT Hyderabad
- IIT Kharagpur

International Institute of Information Technology (IIIT)
Hyderabad - 500032, Telangana, India

31^{st} **March 2015**

Thapar
(Punjabi)

RIT
(Malayalam & Kannada)

DAIICT
(Gujarathi &Marathi)

IITG
(Assamese & Manipuri)

NEHU
(Manipuri)

Tezpur
(Assamese)

IITH
(Telugu & Urdu)

IITKGP
(Bengali & Oriya)

IITK
(Hindi)

IIIT
(Consortium Leader)
(Overall Coordination)

Table of Contents

PART - I

Overview of the Project Work

Development of Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages

General

1 Name of the Project : **Development of Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages**

2 Sanction Letter Reference No. : **11(6)/2011-HCC(TDIL), dated 23-12-2011**

3 Executing Agency :
- : IIIT Hyderabad (Overall Coordination)
- : IIT Kanpur
- : Thapar University Patiala
- : IIT Guwahati
- : Tezpur University
- : North Eastern Hill University (NEHU) Shillong
- : Rajiv Gandhi Institute of Technology (RIT) Kottayam
- : Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT) Gandhinagar
- : IIT Hyderabad
- : IIT Kharagpur

4 Consortium Leader : IIIT Hyderabad
Consortium Head : Prof. B. Yegnanarayana (PI)
Dr. Suryakanth V. Gangashetty (CO-PI)

5 (i) Prinicipal Investigators : Dr. Rajesh Hegde, IIT Kanpur
Prof. R. K. Sharma, Thapar University Patiala
Prof. S. R. Mahadeva Prasanna, IIT Guwahati
Dr. Utpal Sharma, Tezpur University
Dr. L. Joyprakash Singh, NEHU Shillong
Dr. Leena Mary, RIT Kottayam
Dr. Hemant Patil, DA-IICT Gandhinagar
Dr. K. Sri Rama Murty, IIT Hyderabad
Dr. K. Srinivasa Rao, IIT Kharagpur

5 (ii) Co-Investigators Prof. Harish Karnick, IIT Kanpur
Mr. Karun Verma, Thapar University Patiala
Prof. S. Dandapat, IIT Guwahati
Dr. Smriti Kumar Sinha, Tezpur University
Mr. Sushanta Kabir Dutta, NEHU Shillong
Mr. Riyas K. S.& Mr. Anish Babu K. K, RIT Kottayam
Prof. M. V. Joshi, DA-IICT Gandhinagar
Dr. C. Krishna Mohan, IIT Hyderabad
Dr. Pabitra Mitra, IIT Kharagpur

6	Number of Project Staff	:	About 30
7	Total Cost of the Project as approved by DIT		
	i) Original	:	Rs. 492.54 Lakhs
	ii) Revised, if any	:	
8	Project Sanction Date	:	23-12-2011
9	Date of Completion	:	Not Applicable
	i) Original	:	
	ii) Revised, if any	:	
10	Date on which last progress report was submitted	:	28-02-2014

Technical Report

Objective: To develop a prosodically-guided phonetic engine to represent the spoken content and to search speech databases in Indian languages.

Tasks:

- Data collection and manual labeling of speech into phonetic symbols in 12 Indian languages.
- Identifying and marking important prosody events (syllable marking, pitch marking and prosodic breaks marking).
- Development of phonetic engine using prosodic and phonetic information.
- Developing speech-based search engine.

Deliverables:

- **Overall System:**
  - Demonstration of a speech-based search to find data from large speech databases for 12 Indian languages

- **Phonetic Engine Features:**
  - Search 3 hours of speech database in a language using voice input in that language. This will be demonstrated for 12 Indian languages (Hindi, Punjabi, Assamese, Manipuri, Malayalam, Kannada, Gujarati, Marathi, Telugu, Urdu, Bengali, and Odia).
  - Speaker-independent search capabilities.
  - The query consists of 10-30 keywords selected from among the vocabulary of 200-500 words, which are derived from the given databases for that language.
  - Quantitative details of speech databases: 20 hours of speech data in different contexts from at least 20 speakers will be collected for each of the 12 languages.

- **Performance:**
  - From a vocabulary of 200-500, a topic described by about 10-30 words will be used to provide 70-90% relevant result(s) in the top 10 choices. This performance would be measured on 2-3 hours of test data, in each language.
  - Evaluation method indicates precision and recall.

- **Prosody Model:**
  Develop methodology for acquiring prosody knowledge for several (at least 12) Indian languages in 3 different contexts (read speech, lecturing, and conversational speech).

- **Speech Database:**
  Details on Speech data collection and transcription is explained in Chapter 2.

Progress Report : 23-12-2011 to 31-03-2015

Meetings and Workshops:

- December 17, 2010: First preliminary meet
- December 18, 2011: Second preliminary meet
- January 24, 2012: 1st Teleconference
- February 18-20, 2012: Tutorial on phonetic transcription by Prof. Peri Bhaskararao at IIIT Hyderabad (1st Workshop).
- May 11-13, 2012: Intensive Workshop on phonetic transcriptions by Prof. Peri Bhaskararao at IIIT Hyderabad (2nd Workshop).
- October 25-28, 2012: Workshop on prosodically-guided phonetic engine at IIIT Hyderabad (3rd Workshop).
- December 18, 2012: Workshop on phonetic engine and speech-based search engine at IIT Hyderabad (4th Workshop).
- February 1, 2013: 2nd Teleconference
- February 5, 2013: First Project Review of Steering Group meeting at Dept of Electronics & Information Technology, Govt of India New Delhi.
- March 9-10, 2013: Workshop/Meeting at Thapar University Patiala (5th Workshop)
- October 12-13, 2013: Workshop/Meeting at DA-IICT Gandhinagar (6th Workshop)
- March 7-9, 2014: Workshop/Meeting at IIT Kharagpur (7th Workshop)
- September 6-7, 2014: Workshop/Meeting at IIT Guwahati (8th Workshop)
- December 12, 2014: Workshop/Meeting at IIIT Hyderabad (9th Workshop)
- April 30, 2015: Second Project Review of Steering Group meeting at Dept of Electronics & Information Technology, Govt of India New Delhi.

Appendix–A: Summary of the Workshops/Meetings

I. Summary of the 1st workshop conducted by IIIT Hyderabad during February 21–24, 2012

- Introduction to workshop by Prof. B. Yegnanarayana (IIIT Hyderabad)
- International Phonetic Alphabet chart by Prof. Peri Bhaskararao (IIIT Hyderabad)

List of participants (1st workshop)

1. Prof. B. Yegnanarayana, IIIT Hyderabad
2. Prof. Peri Bhaskararao, IIIT Hyderabad
3. Dr. Kishore S Prahallad, IIIT Hyderabad
4. Dr. S. Rajendran, IIIT Hyderabad
5. Shri. Vinay Kumar Mittal, IIIT Hyderabad
6. Dr. Suryakanth V. Gangashetty, IIIT Hyderabad
7. Prof. Rajendra Kumar Sharma, Thapar University Patiala
8. Dr. K. Sri Rama Murty, IIT Hyderabad
9. Dr. S. R. Mahadeva Prasanna, IIT Guwahati
10. Dr Utpal Sharma, Tezpur University
11. Mr. Sushanta Kabir Dutta, NEHU Shillong
12. Dr. L. Joyprakash Singh, NEHU Shillong
13. Dr. Hemant A. Patil, DA-IICT Gandhinagar
14. Dr. K. Sreenivasa Rao, IIT Kharagpur
15. Dr. Debadatta Pati, IIT Kharagpur
16. Dr. R Kumaraswamy, SIT Tumkur
17. Dr. Leena Mary, RIT Kottayam
18. Rupinderdeep Kaur, Thapar University Patiala
19. Ashwini, SIT Tumkur
20. Harish Padaki, IIT Kanpur
21. Deepak, IIT Guwahati
22. Nandakishore, NEHU Shillong
23. Aju Joseph, RIT Kottayam
24. Anish Agustine, RIT Kottayam
25. Maulik C. Madhavi, DA-IICT Gandhinagar
26. Kewal D. Malde, DA-IICT Gandhinagar
27. Sudhamay Maity, IIT Kharagpur
28. Manjunath K.E, IIT Kharagpur
29. Sunil Kumar S. B., IIT Kharagpur
30. Narendra, SIT Tumkur
31. B. Rambabu, IIIT Hyderabad
32. Sudarsana Reddy Kadiri, IIIT Hyderabad
33. Aneeja G., IIIT Hyderabad
34. Karthik Venkat, IIIT Hyderabad
35. P. Gangamohan, IIIT Hyderabad
36. Nivedita Chennupati, IIIT Hyderabad
37. Vishala Pannala, IIIT Hyderabad

38. Naresh Kumar Elluru, IIIT Hyderabad
39. Ravi Shankar Prasad, IIIT Hyderabad
40. Anandaswarup Vadapalli, IIIT Hyderabad
41. Sivanand A, IIIT Hyderabad
42. Bhargav Pulugundla, IIIT Hyderabad
43. Santhosh, IIIT Hyderabad
44. **Sathya Adithya Thati, IIIT Hyderabad**

II. Summary of the 2nd workshop conducted by IIIT Hyderabad during May 11 – 13, 2012

- Introduction to workshop by Prof. B. Yegnanarayana (IIIT Hyderabad)
- International Phonetic Alphabet chart (IPA) by Prof. Peri Bhaskararao (IIIT Hyderabad)
- Prof. Peri Bhaskararao has shown sample phonetic transcriptions for the audio files provided by various consortium members.
- Practice sessions on phonetic transcription was held.
- Data from various consortium members (in different languages) was collected in two modes, recorded in a close room environment with additional EGG setup..
    - Read speech
    - Conversational speech

List of participants (2nd workshop)

1. Prof. B. Yegnanarayana, IIIT Hyderabad
2. Prof. Peri Bhaskararao, IIIT Hyderabad
3. Dr. Kishore S Prahallad, IIIT Hyderabad
4. Dr. S. Rajendran, IIIT Hyderabad
5. Shri. Vinay Kumar Mittal, IIIT Hyderabad
6. Dr. Suryakanth V. Gangashetty, IIIT Hyderabad
7. Dr. Anil Kumar Vuppula, IIIT Hyderabad
8. Dr. N Dhananjaya, IIIT Hyderabad
9. Dr. S. R. Mahadeva Prasanna, IIT Guwahati
10. Mr. Sushanta Kabir Dutta, Co-PI, NEHU Shillong
11. Riyas K.S, (Co-Investigator), RIT Kottayam
12. Dr. Hemant A. Patil, DA-IICT Gandhinagar
13. Dr. K. Sri Rama Murty, IIT Hyderabad
14. Gaurav K Singh, IIT Kanpur
15. Rameshwar Pathak, IIT Kanpur
16. Deepak, IIT Guwahati
17. Biswajit, IIT Guwahati
18. Navanath Saharia, Tezpur University
19. Bhaskar Jyoti Das, Tezpur University
20. Salam Nandakishor, NEHU Shillong
21. Laishram Rahul, NEHU Shillong
22. Aju Joseph, RIT Kottayam
23. Anish Agustine, RIT Kottayam
24. Vachhani Bhavikkumar, DA-IICT Gandhinagar
25. Maulik C. Madhavi, DA-IICT Gandhinagar
26. Kewal D. Malde, DA-IICT Gandhinagar
27. Dipanjan Nandi, IIT Kharagpur
28. Manjunath. K.E, IIT Kharagpur
29. Narendra, SIT Tumkur
30. Rupinderdeep Kaur, Thapar University Patiala
31. Harsimaran Kaur, Thapar University Patiala
32. Mohammad Rafi, IIT Hyderabad

33. Naresh Reddy, IIT Hyderabad
34. Chandan Behra, IIT Hyderabad
35. Anjum Parveen, IIIT Hyderabad
36. Sameena Yasmeen, IIIT Hyderabad
37. Mohammed Younus, IIIT Hyderabad
38. Mohammed Dawood Khan, IIIT Hyderabad
39. Anand Joseph Xavier M., IIIT Hyderabad
40. Gautam Varma Mantena, IIIT Hyderabad
41. B. Rambabu, IIIT Hyderabad
42. Sudarsana Reddy Kadiri, IIIT Hyderabad
43. Aneeja G., IIIT Hyderabad
44. Karthik Venkat, IIIT Hyderabad
45. P. Gangamohan, IIIT Hyderabad
46. Nivedita Chennupati, IIIT Hyderabad
47. Vishala Pannala, IIIT Hyderabad
48. Naresh Kumar Elluru, IIIT Hyderabad
49. Ravi Shankar Prasad, IIIT Hyderabad
50. Anandaswarup Vadapalli, IIIT Hyderabad
51. Sivanand A, IIIT Hyderabad
52. Bhargav Pulugundla, IIIT Hyderabad
53. Santhosh, IIIT Hyderabad
54. Sathya Adithya Thati, IIIT Hyderabad
55. Bajibabu Bollepalli, IIIT Hyderabad
56. Abhijeet Saxena, IIIT Hyderabad
57. Apoorv Reddy, IIIT Hyderabad
58. Basil George, IIIT Hyderabad
59. Sama Vasantha Sai, IIIT Hyderabad
60. Harika Vuppala, IIIT Hyderabad
61. Patha Sreedhar, IIIT Hyderabad

III. Summary of the 3rd workshop conducted by IIIT Hyderabad during 25th - 28th October, 2012

- Review of data collected (Appendix-B)
- Introduction to workshop by Prof. B. Yegnanarayana (IIIT Hyderabad)
    - Syllable marking of spoken data
    - Pitch marking
    - Prosodic break marking
- Data collection
    - Total number of hours of data to be collected: 20 hours
        - 10 hours of read speech, 5 hours of extempore speech and 5 hours of conversational speech have to be collected.
- Data transcription
    - Phonetic transcription of speech data has to be ready before the first PRSG meeting
        - 5 hours of read speech, 2.5 hours of extempore speech and 2.5 hours of conversational speech has to be transcribed.
- Verification
    - Randomly 2 to 3 sentences of transcribed data from all the participating institutions has to be sent to IIIT-H which will be verified by Prof. Peri Bhaskararao (IIIT Hyderabad) .
- Suggestions on evaluation techniques were invited.
- Report on delivery of the data
    - A format has been suggested by Prof. S.R.Mahadeva Prasanna (IIT-G) and will be passed to all consortium members.
- Ideas on prosodically-guided phonetic engine were suggested
    - Phonetic engine based on Place of Articulation (POA) and Manner of Articulation (MOA) was suggested.
    - Tree-based structure implementation was suggested.
- Sample demos were given by IIT-K and IIT-G.
- A basic model of phonetic engine to be made available at the forth coming workshop.
- Discussion on syllable marking of spoken form by Prof. Peri Bhaskararao (IIIT Hyderabad)
    - Definitions of morpheme, word (lexicon), syllable (onset, complex onset, nucleus, coda and complex coda) were explained.
    - The difficulty of identifying syllable boundaries in "natural" conversational speech due to morphophonemic changes was discussed.
    - Discussion on syllable marking of different languages like Telugu and Punjabi language were done.
    - Separate pane for syllabification is suggested in wavesurfer utility.
- Discussion on pitch marking by Prof. Peri Bhaskararao (IIIT Hyderabad)
    - The changes in intonation (pitch patterns) which leads to syntactic changes were discussed.

- Discussion on capturing pitch patterns in spoken form.
- Pitch marking in symbolic form.
- Relative levels of pitch marking were identified, which has to be marked
    - VL – Very Low
    - L – Low
    - H – High
    - VH – Very High
- Issues in marking flatness was discussed and notation f100 to denote the flat pitch 100Hz of the segment (under consideration) was proposed.
- Separate pane for pitch marking is suggested in wavesurfer utility.

- Discussion on prosodic break marking by Prof. Peri Bhaskararao (IIIT Hyderabad)
    - A brief discussion on other types of prosodic labeling methods like ToBI (Tones and Break Indices) was discussed.
    - Prosodic break is obtained due to shift in prosody, but change in pitch patterns is not the only feature responsible for obtaining prosodic breaks.
    - Two types of break indices were discussed
        - Physiological breaks (pauses in spoken form)
        - Prosodical breaks
    - Speech files from various languages were analyzed to show the breaks (b0, b1, b2)
        - b0 – Prosodic break
        - b1, b2– Physiological break
    - In some cases, prosody breaks are identified relatively easily when compared to phonetic information as discussed with Telugu language example.
      (10 such examples from each language need to be identified)
    - Separate pane for break marking is suggested in wavesurfer utility.
    - Transcription to be done in the following format by using wavesurfer utility.

Sl.No.	File Extension	Pane
1	.ph	Phonetic Transcription
2	.sy	Syllable Marking
3	.pt	Pitch Marking
4	.bm	Prosodic Break Marking

- A meta data format has been suggested by Dr. K.Sri Rama Murty (IIT-H) to be followed by all consortium members (Appendix-C)
- Summarization of workshop by Prof. B. Yegnanarayana (IIIT Hyderabad).
- Suggestions and feedback from the participants were collected.

List of participants (3rd workshop)

1. Prof. B. Yegnanarayana, IIIT Hyderabad
2. Prof. Peri Bhaskararao, IIIT Hyderabad
3. Dr. Kishore S Prahallad, IIIT Hyderabad
4. Dr. S. Rajendran, IIIT Hyderabad
5. Shri. Vinay Kumar Mittal, IIIT Hyderabad
6. Dr. Suryakanth V. Gangashetty, IIIT Hyderabad
7. Prof. Rajesh M. Hegde, IIT Kanpur
8. Rameshwar Pathak, IIT Kanpur
9. D. Srinivasulu, IIT Kanpur
10. Dinesh Agnihotri, IIT Kanpur
11. Ms. Rupinderdeep Kaur, Thapar University Patiala
12. Ms. Harsimaran Kaur, Thapar University Patiala
13. Baljinder Baddhan, Thapar University Patiala
14. Prof. S. R. Mahadeva Prasanna, IIT Guwahati
15. Biswajit, IIT Guwahati
16. Prof. Utpal Sharma, Tezpur University
17. Navanath Saharia, Tezpur University
18. Prof. L. Joyprakash Singh, NEHU Shillong
19. Salam Nandakishor, NEHU Shillong
20. Laishram Rahul, NEHU Shillong
21. Prof. Leena Mary, RIT Kottayam
22. Shri Anishbabu K. K. RIT Kottayam
23. Anish Augustine, RIT Kottayam
24. Aju Joseph, RIT Kottayam
25. Prof. Hemant Patil, DA-IICT Gandhinagar
26. Maulik C. Madhavi, DA-IICT Gandhinagar
27. Kewal Dhiraj, DA-IICT Gandhinagar
28. Dr. K. Sri Rama Murty, IIT Hyderabad
29. N. Phanisankar, IIT Hyderabad
30. Mohammand Rafi, IIT Hyderabad
31. Naresh, IIT Hyderabad
32. Chandan Behera, IIT Hyderabad
33. Essa ali khan, IIT Hyderabad
34. Kallol Rout, IIT Hyderabad
35. Sunil Kumar. S. B, IIT Kharagpur
36. Dipanjan Nandi, IIT Kharagpur
37. Manjunath. K. E., IIT Kharagpur
38. Apoorv Chaturvedi, IIT Kharagpur
39. Ravi Kiran, IIT Kharagpur
40. Dr. Debadatta Pati, BCET Balasore
41. Biswajit Sathapathy, BCET Balasore
42. Dr. R Kumaraswamy, SIT Tumkur
43. Narendra K C, SIT Tumkur
44. Shridhar M V, SIT Tumkur

45. Bapu K Banahatti, SIT Tumkur
46. Dr. P. K. Sahu, IIT Bhubaneswar
47. Dr. N.V.L.M Murthy, IIT Bhubaneswar
48. Anand Joseph Xavier M, IIIT Hyderabad
49. B. Rambabu, IIIT Hyderabad
50. Sudarsana Reddy Kadiri, IIIT Hyderabad
51. Aneeja G, IIIT Hyderabad
52. Karthik Venkat, IIIT Hyderabad
53. P. Gangamohan, IIIT Hyderabad
54. Nivedita Chennupati, IIIT Hyderabad
55. Vishala Pannala, IIIT Hyderabad
56. Naresh Kumar Elluru, IIIT Hyderabad
57. Ravi Shankar Prasad, IIIT Hyderabad
58. Anandaswarup Vadapalli, IIIT Hyderabad
59. Ronanki Srikanth, IIIT Hyderabad
60. Sivanand A, IIIT Hyderabad
61. Bhargav Pulugundla, IIIT Hyderabad
62. Santhosh, IIIT Hyderabad
63. Sathya Adithya Thati, IIIT Hyderabad
64. Abhijeet Saxena, IIIT Hyderabad
65. Apoorv Reddy, IIIT Hyderabad
66. Basil George, IIIT Hyderabad
67. S. Vasanth Sai, IIIT Hyderabad
68. Patha Sreedhar, IIIT Hyderabad
69. Padmini Bandi, IIIT Hyderabad
70. G. V. S. Prasad , IIIT Hyderabad
71. Bhanu Teja Nellore, IIIT Hyderabad
72. Sri Harsha Dumpala, IIIT Hyderabad
73. Raghu Ram Nevali, IIIT Hyderabad

IV. Summary of the 4th workshop/meeting conducted at IIT Hyderabad on December 18, 2012

- Development of phonetic engine system
  - Signal level features
  - Acoustic level features (production based)
    (2-3 classes, 10-15 classes, 40-50 classes)
  - Prosody level features
    - Syllable boundary marking
    - Pitch marking
    - Prosodic break marking
    - Prosody models
  - Sound unit level
    - Syllable-like units
    - Phonetic level units
- Searching speech database
  - Target (searching 3-5 hours of speech data in reading mode).
    - Query formation using keywords
    - Query in natural dialog mode (needs word spotting)
  - Other approaches for audio search
- Other issues discussed at the meeting
  - Constitution of internal testing and evaluation committee
  - Suggested dates for PRSG (January 15th, 2013 at IIIT Hyderabad)
  - 5 slides from each group for the PRSG meeting
  - Progress report and UC for the 1st year before December 22, 2012
  - Response to testing and evaluation by CDAC
  - UCs for March & September 2013
  - Discussion on final systems, deliverables and report
  - Next project meeting at Thapar University, Patiala March 2, 2013
  - Discussion in the next meeting will be focused on the systems for phonetic engine and search engine with some demo versions.

List of participants (4th workshop/meeting)

1. Prof. B. Yegnanarayana, IIIT Hyderabad
2. Prof. Peri Bhaskararao, IIIT Hyderabad
3. Dr. Kishore S Prahallad, IIIT Hyderabad
4. Dr. S. Rajendran, IIIT Hyderabad
5. Dr. Anil Kumar Vuppala, IIIT Hyderabad
6. Shri. Vinay Kumar Mittal, IIIT Hyderabad
7. Dr. Suryakanth V. Gangashetty, IIIT Hyderabad
8. Rameshwar Pathak, IIT Kanpur
9. Preeti Singh Chauhan, IIT Kanpur
10. Sukhjeet Kaur, IIT Kanpur
11. Laxmi Pandey, IIT Kanpur

12. Dinesh Agnihotri, IIT Kanpur
13. Dr. Rajendra Kumar Sharma, Thapar University Patiala
14. Rupinderdeep Kaur, Thapar University Patiala
15. Prof. S.R. Mahadeva Prasanna, IIT Guwahati
16. Dr. Utpal Sharma, Tezpur University
17. Navanath Saharia, Tezpur University
18. Dr. L. Joyprakash Singh, NEHU Shillong
19. S. K. Dutta, NEHU Shillong
20. Dr. Leena Mary, RIT Kottayam
21. Riyas K S, RIT Kottayam
22. Anish Babu K K, RIT Kottayam
23. Prof. Hemant A. Patil, DA-IICT Gandhinagar
24. Dr. K. Sri Rama Murty, IIT Hyderabad
25. Dr. C. Krishna Mohan, IIT Hyderabad
26. Dr. K. Sreenivasa Rao, IIT Kharagpur
27. Manjunath. K. E., IIT Kharagpur
28. Sunil Kumar S. B., IIT Kharagpur
29. Dr. R. Kumaraswamy, SIT Tumkur
30. Narendra K C, SIT Tumkur
31. Gautam Varma, IIIT Hyderabad
32. B. Rambabu, IIIT Hyderabad
33. Sudarsana Reddy Kadiri, IIIT Hyderabad
34. Aneeja G, IIIT Hyderabad
35. P. Gangamohan, IIIT Hyderabad
36. Nivedita Chennupati, IIIT Hyderabad
37. Vishala Pannala, IIIT Hyderabad
38. Naresh Kumar Elluru, IIIT Hyderabad
39. Ravi Shankar Prasad, IIIT Hyderabad
40. Anandaswarup Vadapalli, IIIT Hyderabad
41. Sivanand A, IIIT Hyderabad
42. Bhargav Pulugundla, IIIT Hyderabad
43. Santosh K, IIIT Hyderabad
44. Abhijeet Saxena, IIIT Hyderabad
45. Apoorv Reddy, IIIT Hyderabad
46. Basil George, IIIT Hyderabad
47. Patha Sreedhar, IIIT Hyderabad
48. Padmini Bandi, IIIT Hyderabad
49. G. V. S. Prasad, IIIT Hyderabad

V. Summary of the First Project Review of Steering Group meeting at Dept of Electronics & Information Technology, Govt of India New Delhi on February 5, 2013

The Meeting (1) of the Project Review of Steering Group was held on Feb 5, 2013 at Dept of Electronics & Information Technology , Govt of India.

1. **Welcome Address**
   At the beginning, the member convener welcomed the PRSG Chairman and the members in the 1st. PRSG meeting of the ASR in Indian Languages Consortium.
2. **The Chairman** requested the consortium leader to present the progress of the work under the consortium from the date of initiation of the project [i.e from date of issue of the administrative approval -23-12-2011]
3. **Presentation by Consortium Leader**

The Consortium Leader Prof. B. Yegnanarayana presented the progress of the project against the milestones specified in the objectives of the administrative approval.

<table>
<tr><th>Objectives and Milestones for the 1st year</th><th>Progress Reported</th></tr>
<tr><td>

- Collect speech data for a few selected languages for studies on prosody and for development of methodology for searching speech databases in assigned languages.
- Develop algorithms to extract signal level knowledge to incorporate in the PE.
- Explore the prosody and language constraints to improve the performance of the PE

</td><td>

<table>
<tr><th>Language</th><th>Data Collected (in hours)</th><th>Data transcribed phonetically (in hours)</th></tr>
<tr><td>Assamese</td><td>27.40</td><td>14.00</td></tr>
<tr><td>Bengali</td><td>17.42</td><td>09.98</td></tr>
<tr><td>Gujarati</td><td>20.50</td><td>01.75</td></tr>
<tr><td>Hindi</td><td>06.00</td><td>06.00</td></tr>
<tr><td>Kannada</td><td>30.00</td><td>06.00</td></tr>
<tr><td>Malayalam</td><td>21.00</td><td>04.00</td></tr>
<tr><td>Manipuri</td><td>12.00</td><td>06.00</td></tr>
<tr><td>Marathi</td><td>25.50</td><td>03.62</td></tr>
<tr><td>Odiya</td><td>15.00</td><td>05.41</td></tr>
<tr><td>Punjabi</td><td>18.00</td><td>00.42</td></tr>
<tr><td>Telugu</td><td>16.50</td><td>09.50</td></tr>
<tr><td>Urdu</td><td>07.00</td><td>01.00</td></tr>
</table>

The basic algorithm for phonetic based search is developed and being studied for specific requirements for each languages

The work on Syllable marking, Pitch marking , Prosody break marking for each of the above mentioned assigned languages are being carried out by the consortium members responsible for the particular languages.

</td></tr>
</table>

4. Observations of the PRSG

5. PRSG expressed satisfaction about the progress of the project and requested acceleration of the deployment efforts of the Phonetic engines in all assigned languages.

- After development of the Alpha version of the Phonetic engines, the systems may be tested through independent testing and evaluation agency namely C-DAC.
- PRSG also requested to expedite the development process so that, the assigned horizontal and vertical tasks would be completed within the project duration i.e. 23.12.2013.

6. **Recommendations of PRSG.**

The PRSG recommended the release of Next instalment of Grant-in-Aid to IIIT Hyderabad, as per request from the consortium leader after submission of the Utilization Certificates and their acceptance to DEITY. The Meeting ended with a vote of thanks to the Chair.

The list of PRSG Members and the Consortium Members present in the meeting

PRSG Members

Sl No.	Name	Organization	Designation
▪	Dr. P.K.Saxena	Director SAG , DRDO	Chairman
▪	Prof. S.S. Agrawal	Emeritus Scientist , CEERI	Member
▪	Dr. Preeti .S. Rao	IIT Bombay	Member
▪	Prof. Hema Murthy	IIT Madras	Member
▪	Dr. K.Samudravijaya	TIFR , Mumbai	Member
▪	Ms. Swaran Lata	Director &Head, TDIL, DEITY	Member
▪	Dr.Somnath-Chandra	Scientist-E , TDIL, DEITY	Member Convener

Consortium Members

Sl No.	Name	Organization	Designation
1.	Prof. B.Yegnanarayana	IIIT Hyderabad	Consortium Leader
2.	Dr. S.V. Gangashetty	IIIT Hyderabad	Co-PI
3.	Dr. Kishore.S. Prahallad	IIIT Hyderabad	Co-PI
4.	Dr. Rajesh M Hegde	IIT Kanpur	Consortium Member
5.	Prof. Rajendra Kumar Sharma	Thapar University Patiala	Consortium Member
6.	Dr. S R Mahadeva Prasanna	IIT Guwahati	Consortium Member
7.	Dr. Utpal Sharma	Tezpur University	Consortium Member
8.	Dr. L. Joyprakash Singh	North Eastern Hill University (NEHU) , Shillong	Consortium Member
9.	Dr Leena Mary	Rajiv Gandhi Institute of Technology (RIT) Kottayam	Consortium Member
10.	Dr. Hemant A Patil	Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT) Gandhinagar	Consortium Member
11.	Dr. K S R Murty	IIT Hyderabad	Consortium Member
12.	Dr. K Sreenivasa Rao	IIT Kharagpur	Consortium Member

VI. Summary of the 5th workshop/meeting conducted at Thapar University Patiala during February 9th-10th, 2013

The following is the summary of the discussions of two days workshop at Thapar University Patiala:

1. Tasks in the development of phonetic engine system

    1.1 Segmentation: Change in the acoustic features (mostly signal based)
    1.2 Feature stretching: Identify 10-15 features (like voicing, nasality, laterality, etc,) and segments the speech in terms of those features.
    1.3 Phone units: Select/make a subset of classes/labels/categories relevant to each of those
    features in Task 1.2
    1.4 Common phonetic units: Segment the speech based on a subset of common phonetic units
    relevant for each language. This is the direct signal-to-phonetic units, and hence the phonetic engine.

2. Tasks in automating the manual labelling task

    1.1 Phonetic units - IPA chart
    1.2 Syllable boundary marking
    1.3 Pitch accent marking
    1.4 Break index marking

3. Tasks in the development of the search engine system
(Key issues are feature extraction, matching and scoring)

    3.1 Keyword spotting (phone level transcription of the query and the reference and matching)
    3.2 Posteriogram representation and DTW
    3.3 HMM based phone sequences matching
    3.4 Continuous feature vectors
    3.5 Representation by a sequence of a few features (10-15) and then match
    3.6 Vector quantization representation and then matching
    3.7 Spotting phonetic features
    3.8 Combination of some of the above ideas to develop a final system.

4. Other issues
    4.1 Preparing common phonetic symbols from the collected data
    4.2 Identify data for benchmarking: 1 hour of read speech and 1 hour of conversational speech
    for each language, along with query consisting of words/phrases for each case.
    4.3 Prosody definitions:
    -What can be termed as prosody-like?
    -Suprasegmental is only one part

-Voice quality, rhythm, stress, etc.
-Long vowel, segment level, sentence level
-How to acquire prosody information, represent, and exploit it for phonetic engine and search
engine systems.
4.4 System development status and some demos at the next meeting in May 2013, before the PRSG meeting
4.5 Discussion and delivery of search engine system in July workshop at DA-IICT.

List of participants (5th workshop/meeting)

1. Prof. B. Yegnanarayana, IIIT Hyderabad
2. Shri. Vinay Kumar Mittal, IIIT Hyderabad
3. Dr. Hemant A. Patil, DA-IICT Gandhinagar
4. Dr. K. Sreenivasa Rao, IIT Kharagpur
5. Dr. Kishore S Prahallad, IIIT Hyderabad
6. Dr. S. R. Mahadeva Prasanna, IIT Guwahati
7. Dr. K. Sri Rama Murty, IIT Hyderabad
8. Dr. Leena Mary, RIT Kottayam
9. Swati Arora, W3C India
10. Dr. S. Rajendran, IIIT Hyderabad
11. Dr. L. Joyprakash Singh, NEHU Shillong
12. Mr. Sushanta Kabir Dutta, NEHU Shillong
13. Anish Babu K K, RIT Kottaya
14. N Saharia, Tezpur University
15. Deepak, IIT Guwahati
16. Biswajit Dev, IIT Guwahati
17. R Ravi Kiran, IIT Kharagpur
18. Apoorv Chaurvedi, IIT Kharagpur
19. Biswajit Satapathi, IIT Kharagpur
20. Manjunath K E, IIT Kharagpur
21. Baljinder Badham. Tezpur University
22. Kamal Preet Singh, Tezpur University
23. Tarunima Prabhakar, DA-IICT Gandhinagar
24. Mansi Gokhale, DA-IICT Gandhinagar
25. Maulik Madhavi, DA-IICT Gandhinagar
26. Ishtiyaq Husain, IIT Kanpur
27. Rameshwar Pathak, IIT Kanpur
28. Bhavik Vachhani, DA-IICT Gandhinagar
29. Kewal Dheeraj Malde, DA-IICT Gandhinagar
30. Rajesh Hegde, IIT Kanpur
31. Rupinderdeep Kaur, Thapar University Patiala
32. Vishal Kumar, Thapar University Patiala
33. Prof. Rajendra Kumar Sharma, Thapar University Patiala
34. Prof. Smriti K Sinha, Tezpur University
35. Mahinder Singh, Thapar University Patiala
36. Prof. Peri Bhaskararao, IIIT Hyderabad

37. Dr. Suryakanth V. Gangashetty, IIIT Hyderabad
38. Hansi Mean kaur, Thapar University Patiala
39. Dushyant Khurana, Thapar University Patiala

VII. Summary of the 6th workshop/meeting conducted at DA-IICT Gandhinagar during October 12th-13th, 2013

The following is the summary of the discussions of the two days workshop at DA-IICT Gandhinagar:

I. (a) General points by each group
(b) Presentation by each group on audio search and prosody modeling

II. Specifications and issues on audio search

Specifications:
- Refining the system at feature and search level
- Adapting, i.e, improving the performance with usage
- Not diluting the problem, i.e, read speech data (Audio input query: Language specific and mixture of languages)

Issues:
- Input audio query in 5-10 sec data files
- Data - 5 hours, each language in small 5-10 sec data files (for 10 languages), 100-200 keywords.
- Microphone or desktop or webenabled server - local audio search
- Microphone or desktop or not webenabled server
- Query: Subset of keywords - Topic: Study the characteristics of keywords
- Representation of inputs - (a) cluster of phones (b) acoustic features

- Search
    - Approximate string matching - Symbolic
    - DTW variants - Representation and relative emphasis
    - Bag of words - Index (Mapping issues)
    - Prosody constraints in search

III. Phonetic Engine
- Use of acoustic phonetic features and prosody
- Combining different methods
- Evaluation

IV. Prosody labeling
- Syllable marking, pitch accent marking, prosody breaks

V. IIIT Hyderabad to host data from all sites

List of participants (6th workshop/meeting)

1. Prof. B.Yegnanarayana, IIIT Hyderabad
2. Dr. Suryakanth V Gangashetty, IIIT Hyderabad

3. Prof. S. R. Mahadeva Prasanna, IIT Guwahati
4. Prof. K. S. R. Murthy, IIIT Hyderabad
5. Prof. Rajesh Hegde, IIT Kanpur
6. Prof. Leena Mary, RIT Kottayam
7. Prof. L. Joyprakash Singh, NEHU Shillong
8. Prof. S. Rajendran, IIIT Hyderabad
9. Prof. Utpal Sharma, Tezpur Uni., Assam
10. Dr S. K. Sinha, Tezpur Uni., Assam
11. Prof.R. K. Sharma. Thapar University Patiala
12. Dr. K. Sreenivasa Rao, IIT Kharagpur
13. Biswajit Sharma, IIT Guwahati
14. Manjunath K.E, IIT Kharagpur
15. Biswajit Satpathy, IIT Kharagpur
16. Ishtiyaq Husain, IIT Kanpur
17. Rameshwar Pathak, IIT Kanpur
18. Karan Nathwani, IIT Kanpur
19. Deekshitha G, RIT Kottayam
20. Gayathri M. R, RIT Kottayam
21. Shreejith A, RIT Kottayam
22. Shridhara M, V. SIT Tumkur
23. Bapu K Banahatti, SIT,Tumkur
24. Himangshu Sarma, Tezpur University Assam
25. Baljinder, Thapar University Patiala

VIII. Summary of the 7th workshop/meeting conducted at IIT Kharagpur during March 7th-9th, 2014

The following is the summary of the discussions of the three days workshop at IIT Kharagpur:

1. Specific details of the following items were discussed during the concluding session [08-03-2014, 2.30 to 4.00PM]

- Database
- Search engine
- Phonetic engine
- Automatic prosodic transcription

2. Prof. B Yegnanarayana has proposed following things and all the members have agreed

- Form four sub-groups to handle four deliverables (Database, Search engine, Phonetic engine and Automatic phonetic transcription)
- Each sub-group is responsible for assigned deliverables (specification, implementation, documentation and evaluation).
- Sub-groups should work on building systems in all languages.
- Sub-groups to meet in next session and discuss their deliverables.

3. Subgroup details

- **Database**
  1. Dr. S. Rajendran 2. Dr. Hemanth Patil 3. Dr. Utpal Sharma

- **Search engine**
  1. Dr. K S R Murthy 2. Dr. Rajesh Hegde 3. Prof. S R M Prasanna

- **Phonetic engine**
  1. Dr. Suryakanth V G. 2. Prof. S R M Prasanna 3. Dr. Hemanth Patil

- **Automatic prosodic transcription**
  1. Prof. Leena Mary 2. Dr. K. Sreenivasa Rao

4. Suggestion/concerns by consortium members:

- **[Dr. Rajesh Hegde]** concerns over form of delivery of systems: It was decided that for Search engine, Phonetic engine and Automatic prosodic transcription, three totally independent systems will be delivered. Sub-groups are responsible for this system.

- [Prof. B Yegnanarayana] In evaluation of the search engine, in-database and out-of-database key words should be considered.

- [**Dr. Rajesh Hegde**] raised the concern about the time to build systems in all languages.

- [**Dr. K. Sreenivasa Rao**] Manual prosodic transcription.
  Conclusion: Each team should do manual prosodic transcription for 3 modes (in their respective languages). 100 sentences per mode shloud be selected.

- [**Dr. S. Rajendran**] Time stamping of key-words: Key-words used in the search engine should be time stamped. Each team should take this responsibility in their respective languages.

- **[Dr. S. Rajendran]** In database, for read speech text is not available.
  Conclusion: No need to have text for read speech in database.

- **[Prof. B. Yegnanarayana]** Comparison of evaluation results across various modes and languages should be done. Each team should take this responsibility in their respective languages.

  Acknowledge DIT in the literature produced by this project. Database can be shared with anyone.

- **Regarding report:**
  - First draft of the report is ready (done by **Dr. Suryakanth V. Gangashetty**)
  - Report needs to be reviewed.
  - List of accepted papers may be included in the report.

4. Specific details of the following items were discussed during concluding session [08-03-2014, 4.30 to 5.30 PM]

To begin with Prof. B Yegnanarayana gave a informal talk on "Evolution of phonetic engine". Aim of this talk was to throw light on "what should be done, as a follow up to this project".

After this talk, Prof. P. Bhaskara Rao elaborated on the need of acoustic phonetic features.
This was followed by discussion among four sub-groups namely Database group, Search engine group, Phonetic engine group and Prosody labeling group.

- **Action points**

  [Leaders of sub-groups gave a ten minute presentation on the next day meeting about discussion in the sub-group.

- **Evolution of phonetic engine [Prof. B Yegnanarayana]**

  Definition of phonetic engine: Aim of Phonetic Engine (PE) is "to represent what is uttered by a speaker in the form of symbols".

Based on type of symbols used, phonetic engines are classified into five generations.

5. Different approaches for the development of of phonetic engines

- Generation-1 phonetic engine (G1PE)

  G1PE involves representation of speech using parameters (sometimes features) and learning them to identify phoneme. Basically G1PE sees speech as sequence of phoneme -> phoneme to text. G1PE features are not pure acoustic phonetic. Their output is constrained. For example in case of Speech->MFCC->HMM based PE lot of sequential information (may be language) is captured and constraints (lexical, phone sequence itself) the output. Also in G1PE phonemes are used as basic unit and phonemes are language dependent. Hence, there is a need to use unit which is more production oriented.

- **Generation-2 phonetic engine (G2PE)**

  Phoneme is specific to a language. Hence sequence of phonemes is language specific. To overcome this limitation in G2PE, syllable (syl) is used as unit. Syl is a convenient production unit. Syl also imposes production constrains.

  (Features: Something can be seen in acoustic signal: Ex : formant contours.Parameters: blindly extracted from signals using an algorithm: Ex : DFT, LPC, MFCC)

- **Generation-3 phonetic engine (G3PE)**

  Syl could be language dependent. There was need for more production oriented approach. Hence, IPA was used in this project.

- **Generation-4 phonetic engine (G4PE)**

  By transcribing speech using IPA, we might have neglected some production aspects (refer section Acoustic phonetic features [Prof. P. Bhaskara Rao]). Hence, there is a need to have acoustic phonetic description of speech. In G4PE speech will be represented using acoustic phonetic description. Extracting acoustic phonetic description of speech is the challenge.

- **Generation-5 phonetic engine (G5PE)**

  In G5PE speech will be quantized in terms of movement of accumulators, oral cavity description, amount of excitation pressure and etc.

- **Application of acoustic phonetic features [Dr. K S R Murthy]**

  Most of the ASR groups are concentrating on low resource (take model of rich resource language and adapt model to low resource language), zero resource language and multilingual ASR. In traditional ASR Bayesian formulation will be used (probability of the word, given observation seq and model). In case of low and zero resource languages there will be no models and models from high resource languages will be used. In this

context, "word" from low or zero resource language cannot be treated in a traditional way. So word can be considered as sequence of acoustic phonetic description.

- **Acoustic phonetic features [Prof. Peri Bhaskararao]**

  Prof. P. Bhaskararao elaborated on neglected information when a "voiced aspirated plosive" is transcribed. Aim of the talk was to emphasize on Acoustic phonetic (feature) representation of speech. He took four examples of Bengali bh (voiced aspirated) and showed diversity of production features (murmured. modal vowel, voice bar, instant release) and proved these will be lost in the regular transcription.

List of participants (7th workshop/meeting)

1. Prof. B. Yegnanarayana , IIIT Hyderabad
2. Dr. K. Sreenivasa Rao, IIT Kharagpur
3. Mr. Vinay Mittal , IIIT Hyderabad
4. Prof. Peri Bhaskararao, IIIT Hyderabad
5. Dr. Suryakanth V. Gangashetty, IIIT Hyderabad
6. Dr. S. Rajendran , IIIT Hyderabad
7. Dr. K S R Murthy, IIT Hyderabad
8. Dr. Rajesh Hegde , IIT Kanpur
9. Prof. S R Mahadeva Prasanna, IIT Guwahati
10. Dr. Hemanth Patil, DA-IICT Gandhinagar
11. Prof. Leena Mary, RIT Kottayam
12. Prof. R. Kumaraswamy, SIT Tumkur
13. Prof. Rajendra Kumar Sharma, Thapar University Patiala
14. Ms. Rupinderdeep Kaur, Thapar University Patiala
15. Dr. L. Joyprakash Singh , NEHU Shillong
16. Dr. Utpal Sharma, Tezpur University
17. Rameshwar Pathak, IIT Kanpur
18. Biswajit Satpathy, IIT Kharagpur
19. Abhishek Dey , IIT Guwahati
20. Biswajit Sarma, IIT Guwahati
21. Maulik Madhavi, DA-IICT Gandhinagar
22. Jubin James Thennattil, RIT Kottayam
23. Anil P Antony, RIT Kottayam
24. Navanath Saharia, Tezpur University
25. Ishtiyaq Husain, IIT Kanpur
26. Narendra N P, IIT Kharagpur
27. Sunil Kumar S. B., IIT Kharagpur
28. Manjunath K.E, IIT Kharagpur
29. Dipanjan Nandi, IIT Kharagpur
30. Procheta Sen, IIT Kharagpur
31. Parkranth Sarkar, IIT Kharagpur
32. Hari Krishna, IIT Kharagpur
33. Gurunath Reddy, IIT Kharagpur
34. Arijul Haque, IIT Kharagpur
35. Arup Datta, IIT Kharagpur
36. Prasenjit Dhara, IIT Kharagpur

IX. Summary of the 8th workshop/meeting conducted at IIT Guwahati during September 6th-7th, 2014

The meeting is for the purpose of assigning a role to each of consortium members. The following is the summary of the discussions of the two days workshop at IIT Guwahati:

1. Audio Search: [Dr. K. Sri Rama Murty]

- Description of supervised and unsupervised versions of the system with details and results.
- Demonstration of the system with proper user interface.
- Exploring the possibility of incorporating some of the ideas of Phonetic Engine (PE) and prosody for improving the performance.

2. Phonetic Engine: [Prof. S. R. Mahadeva Prasanna]

- Hidden Markov Model (HMM) based training using compressed set of phonetic units derived from manually labelled data
- Graphical User Interface (GUI) for display of the system .
- Language dependent and language independent systems .
- Display of: speech waveform, sequence symbols output, symbols output with confidence threshold, symbols with confidence-based display.
- Overlay of prosody information [**Dr. K. Sreenivasa Rao**]
- Performance evaluation: % Recognition, subjective evaluation of unknown sentences by the user for different displays above .

3. Prosody modelling: [Dr. K. Sreenivasa Rao]

- Display of prosody analysis and evaluation results.
- Integration with PE .

4. Consolidated report:

- Individual reports by 7th October 2014: Dr. S. Rajendran, K. Sri Rama Murty, Prof. S. R. Mahadeva Prasanna, Dr. K. Sreenivasa Rao with Dr. Suryakanth V Gangashetty's help for all.
- Overall report draft by end of October : Prof. B. Yegnanarayana with inputs from Prof. Peri Bhaskararao, also including some future directions.

5. Next proposal:

- Draft by **Prof. B. Yegnanarayana** by the end of September 2014.

List of participants (8th workshop/meeting)

1. Prof. B. Yegnanarayana , IIIT Hyderabad
2. Prof. S R Mahadeva Prasanna, IIT Guwahati
3. Prof. S. Dandapat, IIT Guwahati

4. Dr. Suryakanth V. Gangashetty, IIIT Hyderabad
5. Dr. K Sri Rama Murthy, IIT Hyderabad
6.. Dr. Hemanth Patil, DA-IICT Gandhinagar
7. Dr. K. Sreenivasa Rao, IIT Kharagpur
8. Dr. L. Joyprakash Singh , NEHU Shillong
9. Deepak, IIT Guwahati
10. Vivek C M, IIT Guwahati
11. Biswajit Dev Sarma, IIT Guwahati
12. Abhishek Dey, IIT Guwahati

X. Summary of the 9th workshop/meeting conducted at IIIT Hyderabad on December 12th, 2014

The objective of the meeting is to consolidate the work done and prepare for the following :

(a) Technical report and closure report in the DIT prescribed format
(b) Deliverables
(c) Demonstrations
(d) Utilization Certificates and expenditure statement as per DIT norms
(e) Further action (l**ike next proposal in the direction of Rich Representation**)

The following is the summary of the discussions of the one day workshop at IIIT Hyderabad:

1. **Prof. B Yegnanarayana** - Introduction, review, general format for discussions and writeup and deliveries.
2. **Dr. S Rajendran** - Data collection effort.
3. **Prof. S. R. Mahadeva Prasanna** - Phonetic Engine effort - writeup, demo and delivery.
4. **Dr. K Sreenivasa Rao** - Prosody modelling effort.
5. **Dr. K Sri Rama Murty** - Audio search effort.
6. **Summary, tasks ahead to wind up the project by December 31, 2014.**
   - Writeup consolidation - **Dr. S Rajendran and Dr. Anil Kumar Vuppala**
   - Demos consolidation - **Prof. S. R. Mahadeva Prasanna and Dr. K Sri Rama Murty**
   - Finance and deliverables - **Dr. Suryakanth V Gangashetty and Dr. Anil Kumar Vuppala**

List of participants (9th workshop/meeting)

1. Prof. B. Yegnanarayana, IIIT Hyderabad
2. Dr. S. Rajendran, IIIT Hyderabad
3. Dr. Anil Kumar Vuppala, IIIT Hyderabad
4. Dr. Suryakanth V. Gangashetty, IIIT Hyderabad
5. Prof. Rajendra Kumar Sharma, Thapar University Patiala
6. Dr. K. Sri Rama Murty, IIT Hyderabad
7. Dr. S. R. Mahadeva Prasanna, IIT Guwahati
8. Dr Utpal Sharma, Tezpur University
10. Dr. L. Joyprakash Singh, NEHU Shillong
11. Dr. Hemant A. Patil, DA-IICT Gandhinagar
12. Dr. K. Sreenivasa Rao, IIT Kharagpur
13. Dr. R Kumaraswamy, SIT Tumkur
14. Dr. Leena Mary, RIT Kottayam
15. Gautam Varma Mantena, IIIT Hyderabad
16. Vishala Pannala, IIIT Hyderabad
17. Patha Sreedhar, IIIT Hyderabad
18.. P. Gangamohan, IIIT Hyderabad
19. B. Rambabu, IIIT Hyderabad
20. Sudarsana Reddy Kadiri, IIIT Hyderabad

21. Aneeja G., IIIT Hyderabad
22. Nivedita Chennupati, IIIT Hyderabad
23. Ravi Shankar Prasad, IIIT Hyderabad
24. Anandaswarup Vadapalli, IIIT Hyderabad
25. Sivanand A, IIIT Hyderabad
26. Bhargav Pulugundla, IIIT Hyderabad
27. Santhosh, IIIT Hyderabad

Appendix–B: Data collection and transcription

1. Details of the data collected by IIIT Hyderabad

Language	Read speech (in minutes)	Conversational speech (in minutes)
Assamese	03.39	01.44
Bengali	03.50	05.57
Gujarati	03.19	05.23
Hindi	04.22	03.03
Kannada	03.19	04.03
Malayalam	03.10	05.07
Manipuri	03.32	04.05
Marathi	03.23	05.16
Odia	03.20	05.31
Punjabi	03.31	02.25
Total	**34.05**	**41.14**

2. Data collected and transcribed by the consortium institutes

Language	Source from which data was collected	Data collected (in hours)	Data transcribed phonetically (in hours)
Assamese (1)	Live recording from 4 speakers	13.35	06.00
Assamese (2)	www.newsonair.nic.in, Field recording	09.20	08.00
Bengali	TV, Indoor recording, Field recording	14.10	05.93
Gujarathi	Live recording	14.00	00.30
Hindi	www.newsonair.nic.in, www.youtube.com	02.25	02.25
Kannada	www.newsonair.nic.in, Field recording	35.40	01.50
Malayalam	TV, Field recording	18.00	02.00
Manipuri	www.newsonair.nic.in, Live recording	04.91	04.91
Marathi	Live recording	06.25	00.00
Odia	Indoor recording	10.00	05.00
Punjabi	www.youtube.com	06.10	00.42
Telugu	TV, www.youtube.com	16.50	09.50
Urdu	TV, www.youtube.com	07.00	01.00

Appendix–C: Metadata format for the speech data

1. Metadata file for the speech data

File name: D_INID_GXXXX_LN_MYYYY.info

- Data type :
- Institute ID :
- Gender :
- Speaker ID :
- Recording language :
- Mode :
- File ID :
- Environment :
- Transcribed by :
- Verified by :
- Recording device :
- Close speaking microphone :
- Speaker name :
- Age :
- Mother tongue :
- Home town :
- Home state :
- Place of long term stay :
- Place of stay till age 12 :
- Education :
- Profession :
- Languages :

---

2. Explanation of the fields in the metadata

Field ID	Field Name	Description	A set of possible values/ value type	Remarks
1	D: Data type	Raw data/ Transcribed data	{R, T}	R – Raw data, T – Transcribed data
2	INID	Institute ID	{IIITH, IITH, IITG, IITK, IITKG, RITK, SITT, DAIICT, NEHUS, TEZT, TUPT}	IIITH - International Institute of Information Technology, Hyderabad IITH - Indian Institute of Technology, Hyderabad IITG - Indian Institute of Technology, Guwahati IITK - Indian Institute of Technology, Kanpur IITKG - Indian Institute of Technology, Kharagpur RITK - Rajiv Gandhi Institute of Technology, Kottayam SITT- Sidda Ganga Institute of Technology, Tumkur DAIICT - Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar NEHUS - North Eastern Hill University, Shillong TEZT - Tezpur University, Tezpur TUPT- Thapar University, Patiala
3	G	Gender	{M, F}	M – Male, F - Female
4	XXXX	Speaker ID	Integer	Four digit integer [0001 - 9999]
5	LN	Language of the audio file	{AS, BN, GJ, HN, KN, MN, MR, OD, PN, TE, UR, ML}	AS – Assamese, BN – Bengali, GJ – Gujarati, HN – Hindi, KN – Kannada, MN – Manipuri, MR – Marathi, OD – Odiya, PN – Punjabi, TE – Telugu, UR – Urdu, ML – Malayalam
6	M: Mode	Read speech/ Extempore speech/ Conversational speech	{1,2,3}	1 – Read speech 2 – Extempore speech 3 – Conversational speech
7	YYYY	File number	Integer	Four digit integer [0001 - 9999]
8	Environment	Data collected environment	{Field, Open room, Closed room}	
9	Transcribed by	Name of person who transcribed	String	Max. up to twenty characters

10	Verified by	Name of person who verified the transcription	String	Max. up to twenty characters
11	Recording device	Device used for recording with model number (if any) and manufacturer	{Zoom, Edirols, Set-up box, Others}	Zoom – Zoom model numbers, Edirols – Edirol model numbers, Set-up box – Type of set-up box used, Others - Specify if other than above mentioned devices
12	Close speaking microphone	Whether close speaking microphone setup was used or not	{Y , N}	Y – Yes, N - No
13	Speaker name	Name of the speaker	String	Max. up to twenty characters
14	Age	Age group of the speaker	{10-19, 20-40, 41-60, 61-80 , 81-99}	Age group of 10-19 Years, 20-40 Years, 41-60 Years, 61-80 Years, 81-99 Years
15	Mother tongue	Mother tongue of the speaker	String	Max. up to twelve characters
16	Home town	Home town of the speaker	String	Max. up to twenty characters
17	Home state	Home state of the speaker	String	Max. up to twenty characters
18	Place of long term stay	Long term stay of the speaker	String	Max. up to twenty characters
19	Place of stay till age 12	Place of stay till age 12 of the speaker	String	Max. up to twenty characters
20	Education	Qualification of the speaker	{Below 10th class, 10+2, Graduation, Post Graduation, Other}	
21	Profession	Profession of	String	Max. up to thirty characters

		the speaker		
22	Languages	Languages known by the speaker with read, write, speak options	{Telugu(RWS), English(RS)}	R - Read, W - Write, S - Speak

3. Observations

- The actual file before any processing/modification is defined as a raw file in this document.
- Specifications for audio files:

  Minimum sampling frequency : 16000 Hz
  Minimum bits per sample : 16 bits
  Format : MS WAV format
  Number of channels : Mono/Stereo

  A copy of *ProPEn.conf* will be sent through mail or will be made available for download in the portal shortly.
  Copy this file to the path (in Linux based systems)
  .wavesurfer/x.x/configurations/
  (cp pathwherefileispresent/ProPEn.conf ~/.wavesurfer/x.x/configurations/)
  (x.x is your version of wavesurfer installed in your system e.g.: 1.8)
  Copy this file to the path (in Windows based systems)
  c:\\users\username\.wavesurfer\x.x\configurations\
  (Username is the name of the user in your system)
  (x.x is your version of wavesurfer installed in your system e.g.: 1.8)
  All the data has to be transferred to the Consortium Manager (IIITH) in the following structure.
- Folder structure of data:

  Each institute should send a directory in the name of its Institute ID which consists of folder(s) with Language ID(s) as name of the folder(s) corresponding to languages handled by them. This language folder(s) consists of two different folders namely raw_data and transcribed_data. Each of the raw_data and transcribed_data folders consists of 3 sub-folders. They are Read_speech, Extempore_speech, Conversational_speech corresponding to mode of speech.

  Each speech file in raw_data folder must have only .wav files (*R_INID_GXXXX_LANG_MYYYY.wav*) and corresponding *.info* files.
  Each speech file (*T_INID_GXXXX_LANG_MYYYY.wav*) in transcribed_data folder should be associated with the following files:
  Phonetic Transcription : T_INID_GXXXX_LANG_MYYYY.ph
  Syllable marking : T_INID_GXXXX_LANG_MYYYY.sy
  Pitch Marking : T_INID_GXXXX_LANG_MYYYY.pt
  Prosodic Break Marking: T_INID_GXXXX_LANG_MYYYY.bm
  Metadata file : T_INID_GXXXX_LANG_MYYYY.info

  Transcriptions should be done for continuous sentences. Word level or phoneme level

transcriptions must not be done.
To avoid variations and thus to maintain uniformity across all consortium members, an online metadata acquisition form will be provided shortly.
Applicable only for those members who have done transcriptions in word document

- Open wave file with ProPEn configuration (or) IPA transcription configuration
- Mark (sentence/utterance) boundaries with some common label (say 'u') and silence (with say 'sil') and save the transcription
- Open .ph file (transcription file) in any text editor and replace label (u) with the corresponding phonetic transcription from the word document and save .ph file (This is only one of the ways to convert .doc to .ph for the transcription, no compulsion on following this procedure but one has to submit only .ph files not .doc/.docx/.odt/.txt etc.)

4. Example for the metadata format

File name: R_IIITH_M0125_TE_23458.info

- Data type : Raw
- Institute ID : IIITH
- Gender : M
- Speaker ID : 0125
- Recording language : TE
- Mode : 2
- File ID : 3458
- Environment : Field
- Transcribed by : John
- Verified by : Peter
- Recording device : Zoom
- Close speaking microphone : No
- Speaker name : David
- Age : 20-40
- Mother tongue : Telugu
- Home town : Guntur
- Home state : Andhra Pradesh
- Place of long term stay : Hyderabad

- Place of stay till age 12 : Vijayawada
- Education : Graduation
- Profession : Medical Representative
- Languages : English (RWS), Telugu (RS), Hindi (S)

PART - II

Progress Reports of the Individual Consortium Members

1

IIIT Hyderabad

Progress Report of IIIT Hyderabad

A. General

A.1	Name of the Project	:	**Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages**
	Sanction Letter Reference No.	:	11(6)/2011-HCC(TDIL), Dated 23-12-2011
A.2	Executing Agency	:	IIIT Hyderabad
A.3	Chief Investigator with Designation	:	Prof. B. Yegnanarayana Institute Professor
	Co-Chief Investigators with Designation	:	(1) Dr. Suryakanth V Gangashetty, Assistant Professor (2) Dr. Kishore S. Prahallad, Associate Professor
A.4	Project staff with Qualification (engaged at different periods of time during the project period)	:	(1) Dr. S. Rajendran, Senior Research Officer, Ph.D. (2) Aneeja G., Research Scholar (PhD) (3) Nivedita Chennupati, Research Scholar (PhD) (4) Sathya Adithya Thati, Research Scholar (PhD) (5) P. Gangamohan, Research Scholar, (PhD) (6) Apoorv Reddy, Research Scholar, (MS) (7) Basil George, Research Scholar, (MS) (8) Vishala Pannala, Research Scholar, (MS) (9) B. Rambabu, Senior Research Scholar, (PhD) (10) Karthik Venkat, Research Scholar, (PhD) (11) Sudarsana Reddy K, Research Scholar, (PhD) (12) Patha Sreedhar, Research Scholar, (PhD) (13) Ravi Shankar Prasad, Research Scholar, (MS) (14) Bhanu Teja Nellore, Research Scholar, (MS) (15) Sri Harsha Dumpala, Research Scholar, (MS) (16) Raghu Ram Nevali, Research Scholar, (MS)

A.5 Total Cost of the Project as approved by DIT

(i) Original	:	Rs. 100.05 Lakhs
(ii) Revised, if any	:	

A.6 Project Sanction Date : 23-12-2011

A.7 Date of Completion : Not Applicable

(i) Original	:	
(ii) Revised, if any	:	

A.8 Date on which last progress report was Submitted : 28-02-2014

B. Technical

B.1 Work in Progress (Details are given in technical report in Appendix 1.1)

(a) Database collection in three different modes:

(i) Read speech

(ii) Lecture mode

(iii) Conversational speech

(b) Transcription using IPA chart

(c) Development of prosody models

(d) Development of phonetic engine

(e) Development of speech search application

B.2 Proposed plan of work highlighting the action to be taken to achieve the proposed targets

(a) Report finalization

(b) Database finalization

(c) Code delivery

(d) Finance settlement

C. Project Outcomes

C.1 Papers Published: (See Appendix 1.2)

(i) G. Aneeja and B. Yegnanarayana, "Single Frequency Filtering Approach for Discriminating Speech and Nonspeech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (ASLP)*, vol. 23, no. 4, pp. 705-717, April 2015.

(ii) Vinay Kumar Mittal, B. Yegnanarayana, and Peri Bhaskararao, "Study of the effects of vocal tract constriction on glottal vibration," *The Journal of the Acoustical Society of America (JASA)*, vol. 136, no. 4, pp. 1932-1941, August 2014.

(iii) Anand Joseph Xavier M., Guruprasad Seshadri, and B. Yegnanarayana, "iExtraction of formant bandwidths using properties of group delay functions," *Speech Communication*, vol. 63-64, pp. 70-83, May 2014.

(iv) Gautam Mantena, Sivanand Achanta, and Kishore Prahallad, "Query-by-example spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (ASLP)*, vol. 22, no. 5, pp. 946-955, May 2014.

(v) Gautam Varma Mantena and Kishore S. Prahallad, "Use of articulatory bottle-neck features for query-by-example spoken term detection in low resource scenarios, " *in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP), Florence, Italy*, pp. 7128-7132, May 2014.

(vi) B. George and B. Yegnanarayana, "Unsupervised query-by-example spoken term detection using segment-based bag of acoustic words, " *in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP), Florence, Italy*, pp. 7183-7187, May 2014.

(vii) Ravi Shankar Prasad and B. Yegnanarayana, "Acoustic segmentation of speech using zero time liftering (ZTL)," in *Proc. Interspeech, Lyon, France*, pp. 2292-2296, August 2013.

(viii) Apoorv Reddy, Nivedita Chennupati and B. Yegnanarayana, "Syllable nuclei detection using perceptually significant features," in *Proc. Interspeech, Lyon, France*, pp. 963-967, August 2013

(ix) Dhananjaya N., B. Yegnanarayana, and Peri Bhaskararao, "Acoustic analysis of trill sounds," *The Journal of the Acoustical Society of America (JASA)*, vol. 131, no. 4, pp. 3141-3152, April 2012.

C.2 Development of Database

(i) Data was collected from participants of workshops held in IIIT Hyderabad

C.3 Tools and Systems Developed

(i) Templates for data transcription and prosody labeling was developed

(ii) Phonetic engine template for syllable labeling

(iii) Audio search template

(iv) Acoustic phonetic labeling template

(v) Prosody labeling

Appendix 1.1

Detailed Technical Report of IIIT Hyderabad

1.1 Responsibilities

IIIT Hyderabad is responsible for overall coordination of the project. It has planned several meetings and workshops as listed in the overview of the project. In addition the following tasks are being carried out.

1.2 Database collection and transcription

- Data was collected from the participants in the workshop in their respective languages in different contexts.
- This data will be transcribed for use in the development of prosody models and also the phonetic engine
- The templates for data collection and labeling were designed and distributed to all the consortium partners

1.3 Prosody Knowledge

The guidelines for acquiring the prosody knowledge were evolved for the following subtasks:

(a) Syllabification of spoken data

(b) Pitch marking

(c) Marking prosody breaks

1.4 Development of Phonetic Engine

Several versions of phonetic engine are under development

(a) A system for speech signal to syllable transcription using the conventional HTK toolkit with some modifications taking into account constraints only at syllable level (with no language constraints) was developed. It gives a syllable level accuracy of about 50 %.

(b) Phonetic engine system using acoustic level features is being developed using less than 10 acoustic features.

1.5 Development of Speech-based Search Engine

Three different approaches are being explored for keyword spotting in read speech

(a) Dynamic time warping (DTW) based system

(b) System based on phonetic/phonemic sequence representation

(c) System based on acoustic feature representation

1.6 Summary of the Work

IIIT Hyderabad is mainly engaged in coordinating the activities of different group working in different languages. In particular, workshops were organized to expose the participating groups to the concepts of phonetic labeling and prosody labeling. Also the guidelines for phonetic engine system and search system are being evolved and communicated to participating members.

In addition, small amount of data was collected in different languages and in different contexts for phonetic labeling and prosody labeling by this group. These results will be used for the development of phonetic engine modules and search engine. Versions of phonetic engine and search engine are developed.

Appendix 1.2

List of Publications

Single Frequency Filtering Approach for Discriminating Speech and Nonspeech

G. Aneeja and B. Yegnanarayana, *Fellow, IEEE*

***Abstract*—In this paper, a signal processing approach is proposed for speech/nonspeech discrimination. The approach is based on single frequency filtering (SFF), where the amplitude envelope of the signal is obtained at each frequency with high temporal and spectral resolution. This high resolution property helps to exploit the resulting high signal-to-noise ratio (SNR) regions in time and frequency. The variance of the spectral information across frequency is higher for speech and lower for many types of noises. The mean and variance of the noise-compensated weighted envelopes are computed across frequency at each time instant. Decision logic is applied to the feature derived from the mean and variance values on varieties of degradations, including NTIMIT, CTIMIT and distance speech, besides degradation due to standard noise types. In all cases, the proposed method gives significantly better performance than the standard Adaptive Multi-rate VAD2 (AMR2) method. AMR2 method is chosen for comparison, as the method adapts itself for different degradations, and is seen to give good performance over different SNR situations. The proposed method does not use training data to derive the characteristics of speech or noise, nor makes any assumption on the nonspeech beginning. The SFF method appears promising in other applications of speech processing, such as pitch extraction and speech enhancement.**

***Index Terms*—Single frequency filtering (SFF), spectral variance, speech/nonspeech discrimination, temporal variance, voice activity detection (VAD), weighted component envelope.**

I. Introduction

THE objective of voice activity detection (VAD) is to determine regions of speech in the acoustic signal, even when the signal is corrupted by additive or other types of degradations. VAD is an essential first step for development of speech systems such as speech and speaker recognition. Human listeners are able to distinguish speech and nonspeech regions by interpreting the signal in terms of speech characteristics, as well as the context. If a machine has to discriminate these two regions, it has to depend only on the characteristics of speech and degradation. It is difficult to make a machine use the accumulated knowledge of a human listener for this purpose.

Robustness of a VAD algorithm depends on the type of degradation, the features extracted from the signal and the models used to discriminate speech and nonspeech regions. The acoustic features are usually based on the signal energy in different frequency bands, which includes standard melfrequency cepstral coefficients (MFCC's) [1]. Features based on speech characteristics such as voicing and dynamic spectral characteristics have also been explored [2], [3]. In [2], the phase of the Fourier Transform is averaged over a window to compensate for phase wrapping, and then processed over mel-frequency bands. The phase information gives performance similar to MFCCs even in the cases of degradation. But combination of MFCCs and phase information seems to have improved the performance. Some attempts have been made to explore features in the excitation component of speech signal [4]. Features of the discrete wavelet transform and Teager energy operator have also been proposed for VAD with good results [5], [6]. Characteristics of speech and noise can be captured well if the samples are collected over long ($>$ 1 sec) durations, as some of the studies below indicate. For example, the long-term divergence measure (LTDM) measures the spectral divergence between speech and noise over longer duration [7]. The LTDM measure is calculated as the ratio of the long-term spectral energies of speech and noise over different frequency bands. More recently long-term spectral variability has been suggested for VAD [8]. The long-term feature is the variance across frequency of the entropy computed over 300 msec of speech at each frequency. It was shown to be robust at low signal-to-noise ratio (SNR) conditions for a variety of noise degradations. The long-term signal variability (LTSV) was extended to multi-band long-term signal variability to accommodate multiple spectral resolutions [9]. The long-term spectral variability feature together with contextual, discriminative and spectral cues was shown to give further improvement in performance of VAD [10]. New features like Multi-Resolution cochleagram (MRCG) along with boosted Deep Neural Networks (bDNNs) have been proposed recently for VAD, which are shown to outperform the state-of-the-art VADs even at low SNRs, for babble and factory noises [11], [12]. The MRCG feature is derived using features at multiple spectrotemporal resolutions [11] and the bDNN uses aggregate of predictions of multiple weak classifiers [12].

In [13], a low variance for spectral estimate is assumed for noise, and large amount of data is used for training. But low variance criterion for noise may not be applicable for machine gun noise and some other non-stationary noises, including distant speech. The method proposed in [13] assumes a nonspeech beginning to estimate the noise statistics. Other models are also considered for speech and nonspeech discrimination, which include artificial neural networks (ANNs) [14], Gaussian mixture models (GMMs) [15], and deep belief networks (DBNs) [16].

Manuscript received July 19, 2014; revised November 14, 2014; accepted January 31, 2015. Date of publication February 13, 2015; date of current version March 06, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yunxin Zhao.

The authors are with the International Institute of Information Technology, Hyderabad 500 032, India (e-mail: aneeja.g@research.iiit.ac.in; yegna@iiit.ac.in).

Digital Object Identifier 10.1109/TASLP.2015.2404035

Several attempts have been made to improve the performance of VAD, by exploiting the statistics of speech and noise characteristics [17]. One such method is the statistical model-based VAD, and its refinements proposed in [18]. Statistical methods work well if labelled training data for speech and nonspeech in different noise conditions are available for training the models. These are called supervised learning systems [19]. In some cases, the noise model derived from training data is used for initialization process. These methods are called semi-supervised learning [17]. Methods based on universal models of speech, without assuming any specific type of noise, are also proposed. In [20], non-negative matrix factorization (NNMF) approach is used to develop universal speech model. In practice, it is preferable to develop a VAD algorithm that can operate without any training data, i.e., unsupervised learning.

Most of the VAD algorithms are tested on data with simulated degradation, either by adding noise or by passing the clean signal through a degrading channel. This is necessary to evaluate new methods in comparison with known/existing methods. Very few attempts have been made to assess the performance of a VAD algorithm with data collected in practical environments. The degradations in such environments may not fit into any standard model. Moreover, it is difficult to obtain ground truth in practice to evaluate the VAD methods. In general, the characteristics of the environment in which speech signal is produced vary, and hence are not predictable to model. The only option available is to develop VAD algorithm by exploiting the characteristics of speech that may be present even in the degraded signal. For this, the features of excitation source and dynamic vocal tract system need to be explored for robustness against degradation. Also, it is necessary to develop methods to extract those features from degraded signals.

In this paper, a signal processing approach is proposed to highlight the features of speech even in degraded signal. The method extracts the temporal variation of signal energy at each frequency. The characteristics of speech signal (due to correlations among speech samples) at each frequency are distinctly different from the characteristics of noise (due to uncorrelatedness in noise samples in many cases) at each frequency. The SNR of the speech signal is high at some frequencies, compared to noise. The high SNR property of speech at several single frequencies is exploited. Since the method is based on extracting energy at a single frequency, it is called single frequency filtering (SFF) method. Note that single frequency information can also be derived by computing the discrete Fourier transform (DFT) over a block of data at every sampling instant. Other methods of deriving similar information include gammatone filters [21]. The temporal variation of signal energy at each frequency is processed further to compensate for the effects of noise in that band by determining a weighting factor for each band. The mean and variance of the weighted signal energy across frequency at each sampling instant are used to derive a parameter contour as a function of time, to discriminate between speech and nonspeech regions. An adaptive threshold is derived from the parameter contour for each utterance, followed by a decision logic based on the features of speech and noise in the given utterance. The method is tested using simulated degradations on speech signals, and also using speech signals collected in practical environments. Since the method exploits the properties of the speech signal, it is not necessary to have training data of speech and nonspeech signals to build models. The present approach does not rely on the appended silence/noise regions to estimate the noise characteristics.

Many studies in literature compare VAD algorithms with the Adaptive Multi-Rate (AMR) method [22]. The comparison is done mainly at the score level. To have a fair comparison with the AMR method, the VAD algorithms should consider the following other factors of the AMR method into account:

- Adaptability: AMR method is adaptable to various types of noise, SNRs and environments.
- No prior information: It does not require training data or any other prior information about the type of noise.
- Automatic threshold: The threshold estimation does not require nonspeech beginning, and also does not use data for training of statistical models.

In Section II, speech data collected in different types of degradation is described. Section III discusses the basis for the proposed single frequency filtering (SFF) method for processing the signals. Section IV gives the development of the proposed VAD algorithm. Section V gives results of evaluation of the SFF-based method of VAD in comparison with the AMR2 method for different types of degradations. This section also includes a discussion on relative performance of SFF, DFT and gammatone filtering methods of deriving information in different frequency bands. Section VI gives a summary, and indicates how the proposed SFF method can be exploited for other speech processing applications.

II. Different Types of Degradation

In this section different speech and noise databases and their characteristics are discussed to indicate the variety of degradations considered for evaluation of the proposed VAD algorithm. Note that, although some of the data was collected at 16 kHz sampling rate and other data at 8 kHz sampling rate, the frequencies in the range 300 - 4000 Hz are considered in both the cases as explained in Section IV-A.

A. Adding Degradation at Different SNRs to Clean Speech Signal

The TIMIT test corpus is used for evaluation [23]. The sampling rate is 16 kHz. A VAD algorithm should ideally accept speech and also reject nonspeech. In a situation where there is more duration of speech than nonspeech, then if the algorithm has a higher speech acceptance, then the algorithm shows better performance even if the performance of nonspeech rejection is poor. A similar situation of better performance would arise for longer duration of nonspeech, with higher nonspeech detection rate and lower speech detection rate of the algorithm. To overcome this problem, each TIMIT utterance is appended with 2 sec of silence at the beginning and end of the utterance as in [8]. Various samples of the thirteen types of noises from NOISEX-92 database [24] are added to the clean TIMIT speech signal at SNRs of -10 dB and 5 dB, to create degraded speech signals. The TIMIT data provides boundaries of the phone labels, which are generated automatically and are then hand corrected by experienced acoustic phoneticians. Hence these boundaries

are used as ground truth for comparing the results of the proposed VAD algorithm on the noisy speech data. The silence and pause labels are considered as nonspeech.

Most VAD algorithms use post processing techniques like hangover scheme. The hangover scheme is used to reduce the risk of lower energy regions of speech at the ends of speech regions being falsely rejected [13]. This is based on the assumption that speech frames are highly correlated in time [13], [17]. In hangover schemes decisions at the frame level are smoothed by considering sequence of frames to arrive at a final decision. Hangover schemes are applied to the VAD algorithm after the initial VAD decision. In some regions, the features of speech might not be evident even in clean speech, although those regions are labelled as speech in the database. The ground truth given in TIMIT database may not be a perfect reference for comparing results of any VAD algorithm. This may be due to mismatch between the perceptual evidence and speech data in manual labelling. Hence the accuracy will not be 100% even in the case of clean speech.

B. Telephone Channel Database

NTIMIT (Network TIMIT) database [25] was collected by transmitting TIMIT data over telephone network. Speech utterances are transmitted from a laboratory to a central office and then back from the central office to the laboratory, thus creating a loopback telephone path from laboratory to a large number of central offices. These central offices were geographically distributed to simulate different telephone network conditions. Half of the TIMIT database was sent over local telephone paths, while the other half was transmitted over long distance paths. All recordings were done in an acoustically isolated room. The NTIMIT test corpus is used for VAD evaluation. The sampling rate is 16 kHz. In the NTIMIT case, 2 sec silence segments are not appended to the data, as this kind of degradation can not be simulated in the appended regions. The ground truth for the NTIMIT is same as for the TIMIT data.

C. Cellphone Channel Database

The CTIMIT read speech corpus [26] was designed to provide a large phonetically-labelled database for use in the design and evaluation of speech processing systems operating in diverse, often hostile, cellular telephone environments. CTIMIT was generated by transmitting and redigitizing 3367 of the 6300 original TIMIT utterances over cellular telephone channels from a specially equipped van, in a variety of driving conditions, traffic conditions and cell sites in southern New Hampshire and Massachussetts. The recorded data was played in the van over a loudspeaker and cellular handset combination. Each received call was digitized at 8 kHz, segmented and time-aligned with the original TIMIT utterances. The ground truth of TIMIT labels can be used here also. CTIMIT test corpus is used for VAD evaluation [26]. Note that here also the 2 sec silence segments are not appended to the data, as in the case of NTIMIT database.

D. Distant Speech

The differences between the characteristics of speech signal collected by a distant microphone (DM) and that collected by a close-speaking microphone (CM) are as follows: (a) The effects of radiation at far-field are different from those at the near-field. (b) The SNR is lower in the DM speech signal due to additive background noise. (c) The reverberant component in the DM speech signal is also significant, due to reflections, diffuse sound and reduction in amplitude of the direct sound. (d) The DM speech signal may also be affected due to interference from speech of other speakers present in the room. Hence, the acoustic features derived from the DM speech signal are not same as those derived from the corresponding CM speech signal.

Speech signals from SPEECON database are used for evaluation of the VAD algorithm for distant speech [27]. The signals were collected in three different cases, namely, car interior, office and living rooms (denoted by public). The signals were collected simultaneously using a close-speaking microphone (a microphone placed just below the chin of the speaker), and microphones placed at distances of 1 meter, 2 meters and 3 meters from the speaker. These four cases are denoted by C0, C1, C2 and C3, respectively. Each case has 1020 utterances. Speech signals collected in the office environment are affected by noises generated by computer fans and air-conditioning. Speech signals collected in living rooms are affected by babble noise and music (due to radio or television sets). Reverberation is present mostly in the office and living room environments. The estimated reverberation time in these environments varied from 250 msec to 1.2 sec. The average SNR measured at the close speaking microphone (C0) was around 30 dB, while that measured at distances of 2 meters to 3 meters was in the range $0-5$ dB. The database consists of speech signals collected from 30 male and 30 female speakers. For each speaker, 17 utterances were recorded, resulting in about one minute of speech data per speaker. People were asked to record free spontaneous items, elicited spontaneous items, read speech and core words. A manual voiced-unvoiced-nonspeech labels are marked for every 1 msec in the SPEECON database for C0 case. Since speech at all the distances are simultaneously collected, the same labels are used for the data at all distances. The manual labels (voiced-unvoiced labels for speech and nonspeech label for the rest) form the ground truth for the data at all distances. The sampling rate is 16 kHz. Since the utterances of each speaker are from different environments, it is not possible to build statistical models with this kind of data. No silence data is appended in this case also.

III. Basis for Single Frequency Filtering Approach

Speech signal has dependencies both along time and along frequency. This results in signal to noise power ratio to be a function of time as well as a function of frequency. For an ideal noise of a given total power, the power gets divided equally over frequency, whereas for a signal, the power is distributed nonuniformly across frequency. Thus $\frac{S^2(f)}{N^2(f)}$ is higher in some frequencies and lower in some other frequency regions, where $S(f)$ and $N(f)$ are signal and noise amplitudes as a function of frequency. This gives a much higher value for the average of $\frac{S^2(f)}{N^2(f)}$ over a frequency range, compared to the ratio of total signal power to total noise power over the entire frequency range.

Let

$$\alpha = \int_{f_0}^{f_L} \frac{S^2(f)}{N^2(f)} df, \tag{1}$$

$$\beta = \sum_{i=0}^{L-1} \frac{\int_{f_i}^{f_{i+1}} S^2(f) df}{\int_{f_i}^{f_{i+1}} N^2(f) df}, \tag{2}$$

and

$$\gamma = \frac{\int_{f_0}^{f_L} S^2(f) df}{\int_{f_0}^{f_L} N^2(f) df}, \tag{3}$$

where $(f_i - f_{i+1})$ is the $(i+1)$th interval of the L nonoverlapping frequency bands, and $i = 0, 1, \ldots, L-1$. The following inequality holds good.

$$\alpha \geq \beta \geq \gamma. \tag{4}$$

The $S(f)$ and $N(f)$ are computed for degraded speech utterance and for noise using 512-point DFT of Hann windowed segments of size 20 msec for *every sample shift* using $L = 16$. In Table I, the mean values $\bar{\alpha}$, $\bar{\beta}$, $\bar{\gamma}$ of α, β and γ respectively, computed over the entire utterance are given. It is clear that $\bar{\alpha} \geq \bar{\beta} \geq \bar{\gamma}$ for different types of noises. In the case of uniform noise, (eg white), the values of $\bar{\alpha}$, $\bar{\beta}$, $\bar{\gamma}$ are lower than the values for the nonstationary noises (eg volvo and machine gun). In the case of some nonstationary noises, the floor value is low at some frequencies which makes the denominator $N(f)$ small. With small values of the denominator, the ratios of α, β, γ are relatively higher as observed in Table I from the values of $\bar{\alpha}$, $\bar{\beta}$, $\bar{\gamma}$ for volvo and machine gun noises. It is also interesting to note that for nonuniformly distributed noises, such as machine gun, f16 and volvo, the $\bar{\alpha}$ and $\bar{\beta}$ values are much higher than for the more uniformly distributed noises, such as white, pink and buccaneer2, whereas the corresponding $\bar{\gamma}$ values are low in all cases. This is due to regions having high $\frac{S(f)}{N(f)}$ in the time and frequency domains for nonuniformly distributed noises.

The signal and noise power as a function of frequency can be computed using either by block processing as in the DFT, or by filtering through SFF, as described in the next section. Table II shows that the inequality (4) holds good for SFF approach also. Both the DFT and SFF based approaches are expected to give similar results. The SFF approach is used here, as it may avoid some effects due to block processing. Also, the computation of SFF is faster compared to the computation of DFT at each sampling instant.

TABLE I
VALUES OF $\bar{\alpha}$, $\bar{\beta}$, $\bar{\gamma}$ FOR SPEECH SIGNAL DEGRADED AT -10 dB SNR USING DFT APPROACH

NOISE	$\bar{\alpha}$	$\bar{\beta}$	$\bar{\gamma}$
white	2.289	1.1224	1.103
babble	237.4061	8.518	1.1481
volvo	3698.2985	233.7515	1.4089
leopard	1599.8217	35.0032	1.143
buccaneer1	277.1179	1.1356	1.1166
buccaneer2	5.4785	1.1299	1.0975
pink	2.2999	1.1034	1.1105
hfchannel	132.6712	1.8899	1.1094
m109	79.2124	2.4393	1.1294
f16	34927.2057	1.3335	1.1098
factory1	406.8931	1.2621	1.1199
factory2	66.5143	3.626	1.1703
machine gun	186034.6397	12056.4657	71.9059

TABLE II
VALUES OF $\bar{\alpha}$, $\bar{\beta}$, $\bar{\gamma}$ FOR SPEECH SIGNAL DEGRADED AT -10 dB SNR USING SFF APPROACH

NOISE	$\bar{\alpha}$	$\bar{\beta}$	$\bar{\gamma}$
white	3.9853	1.1317	1.1034
babble	804.0397	3.992	1.1596
volvo	299.3464	44.2498	1.4496
leopard	117.9238	10.9565	1.156
buccaneer1	72.916	1.1395	1.1184
buccaneer2	3.1214	1.134	1.0978
pink	4.9493	1.1131	1.1112
hfchannel	17.8483	1.6226	1.1113
m109	59.2573	2.0275	1.1414
f16	18.2013	1.2939	1.1126
factory1	4.9478	1.2516	1.1235
factory2	21.4015	2.3495	1.1777
machine gun	10642.4183	753.7107	69.1977

IV. PROPOSED VAD ALGORITHM

A. Envelope of Speech Signal at Each Frequency

The discrete-time speech signal $s(n)$ is differenced, and the differenced signal is denoted by $x(n) = s(n) - s(n-1)$. The sampling frequency is f_s. The signal $x(n)$ is multiplied by a complex sinusoid of a given normalized frequency $\bar{\omega}_k$. The resulting operation in the time domain is given by

$$x_k(n) = x(n) e^{j\bar{\omega}_k n}, \tag{5}$$

where

$$\bar{\omega}_k = \frac{2\pi \bar{f}_k}{f_s}. \tag{6}$$

Since we multiplied $x(n)$ by $e^{j\bar{\omega}_k n}$, the resulting spectrum of $x_k(n)$ is a shifted spectrum of $x(n)$. That is,

$$X_k(\omega) = X(\omega - \bar{\omega}_k), \tag{7}$$

where $X_k(\omega)$ and $X(\omega)$ are spectra of $x_k(n)$ and $x(n)$, respectively.

The signal $x_k(n)$ is passed through a single-pole filter, whose transfer function is given by

$$H(z) = \frac{1}{1 + rz^{-1}}. \tag{8}$$

The single-pole filter has a pole on the real axis at a distance of r from the origin. The the location of the root is at $z = -r$ in the z-plane, which corresponds to half the sampling frequency i.e., $f_s/2$. The output $y_k(n)$ of the filter is given by

$$y_k(n) = -r y_k(n-1) + x_k(n). \tag{9}$$

The envelope of the signal $y_k(n)$ is given by

$$e_k(n) = \sqrt{{y_{kr}}^2(n) + {y_{ki}}^2(n)}, \tag{10}$$

where $y_{kr}(n)$ and $y_{ki}(n)$ are the real and imaginary components of $y_k(n)$. Since the filtering of $x_k(n)$ is done at $f_s/2$, the above envelope $e_k(n)$ corresponds to the envelope of the signal $x_k(n)$ filtered at a desired frequency of

$$f_k = \frac{f_s}{2} - \bar{f}_k. \tag{11}$$

The above method of estimating the envelope of the component at a frequency f_k is termed as single frequency filtering (SFF) approach. The choice of the filter with a pole at $z = -r$ for estimating the envelopes of the filtered signals is likely to be more accurate, as the envelopes are computed at the highest frequency ($f_s/2$) possible. Also, choosing a filter at a fixed frequency for any desired frequency f_k avoids scaling effects due to different gains of the filters at different frequencies. If the pole is chosen on the unit circle, i.e., $z = r = -1$, it may result in the filtered output becoming unstable. The stability of the filter is ensured by pushing the pole slightly inside the unit circle. Hence r is chosen as 0.99.

In this study, the envelope is computed at every 20 Hz in the range 300 Hz to 4000 Hz as a function of time. The frequency range 300 - 4000 Hz is chosen, as it covers the useful spectral band of speech. Thus we have envelopes for 185 frequencies as a function of time. In principle, the envelope can be computed at any desired frequency.

B. Weighted Component Envelopes of Speech Signal

Since speech signal has large dynamic range in the frequency domain, the signal may have high power at some frequencies at each instant. At those frequencies the SNR will be higher, as the noise power is likely to be less due to more uniform distribution of the power. Even for noises with nonuniform distribution of power, the lower correlations of noise samples result in a lower dynamic range in the spread of noise power across frequencies, compared to speech. Note that the spectral dynamic range gives an indication of the correlation of the samples in the time domain.

The noise power creates a floor for the envelope at each frequency, and the floor level depends on the power distribution of noise across frequency. The floor is more uniform across time if the noise is nearly stationary. Even if the noise is nonstationary, it is relatively stationary over larger intervals of time than in speech. In such cases, the floor level can be computed over long time intervals at each frequency, if needed.

To compensate for the effect of noise, a weight value at each frequency is computed using the floor value. For each utterance, the mean (μ_k) of the lower 20% of the values of the envelope at each frequency f_k is used to compute the normalized weight value w_k at that frequency. The choice of 20% of the values is based on the assumption that there is at least 20% of silence in the speech utterance. The normalized weight value at each frequency is given by

$$w_k = \frac{\frac{1}{\mu_k}}{\sum_{l=1}^{N} \frac{1}{\mu_l}}, \tag{12}$$

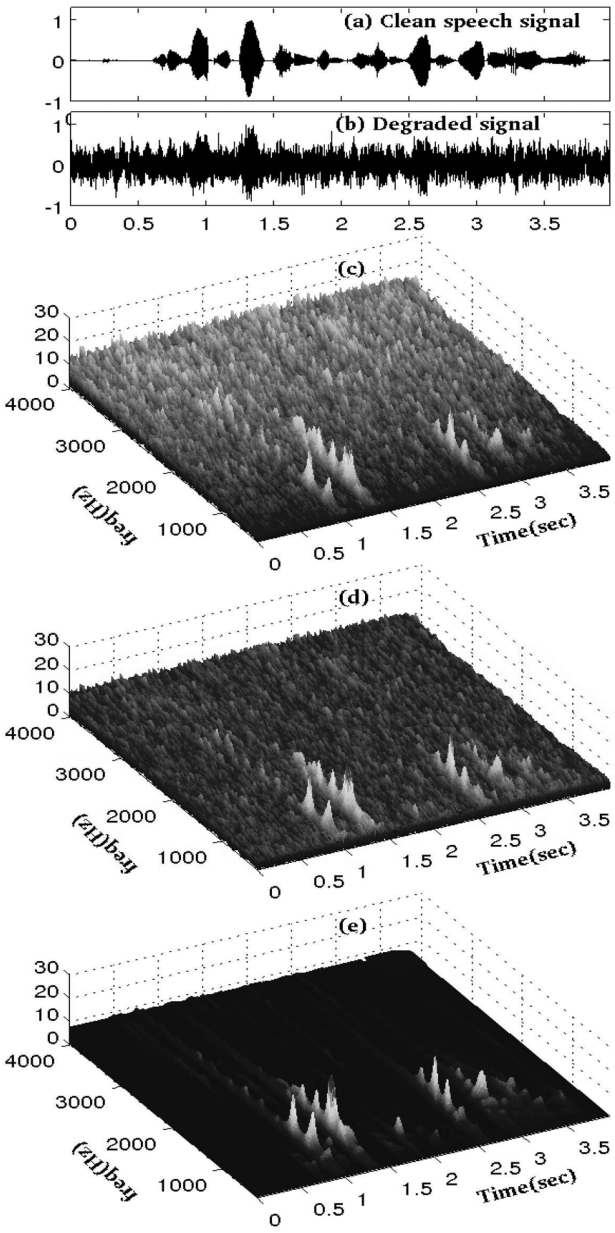


Fig. 1. (a) Clean speech signal. (b) Speech signal corrupted by pink noise at -10 dB SNR. (c) Envelopes as a function of time. (d) Corresponding weighted envelopes. (e) Envelopes as a function of time for clean speech shown in (a).

where N is the number of channels. The envelope $e_k(n)$ at each frequency f_k is multiplied with the weight value w_k to compensate for the noise level at that frequency. The resulting envelope is termed as weighted component envelope. Note that by this weighting, the envelope at each frequency is divided by the estimate of the noise floor (μ_k). Fig. 1 shows the envelopes and the corresponding weighted envelopes at different frequencies for a speech signal degraded by pink noise at -10 dB SNR, along with the envelopes for clean speech. It is observed that features of speech are reflected better in the weighted envelopes (Fig. 1(d)), as the weighting reduces the effects of noise. The envelopes are scaled to the same value for comparison.

A small amount of white noise (at 100 dB SNR) is added to all the signals (after appending with zeros in the case of TIMIT utterances) to ensure that the floor value is not zero. For the

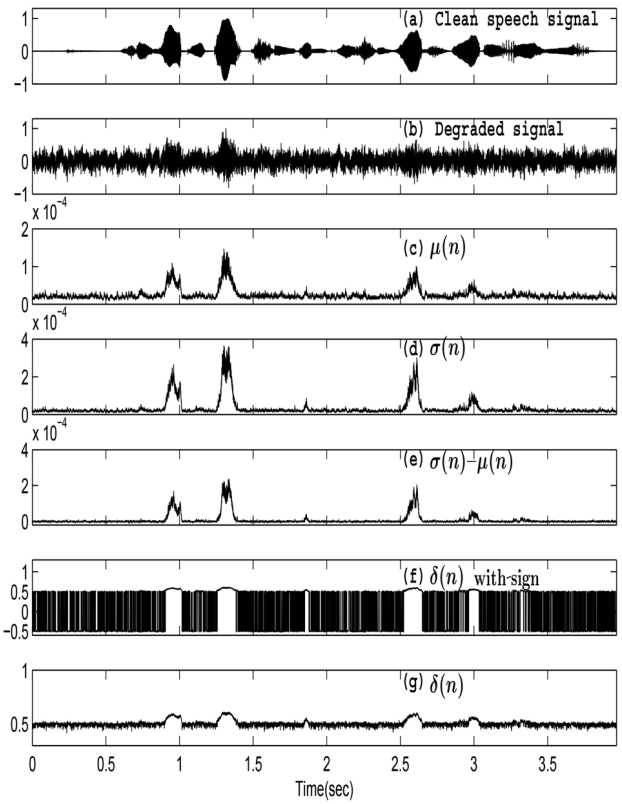


Fig. 2. (a) Clean speech signal. (b) Speech signal corrupted by pink noise at -10 dB SNR. (c) $\mu(n)$. (d) $\sigma(n)$. (e) $\sigma(n) - \mu(n)$. (f) $\delta(n)$ along with sign. (g) $\delta(n)$.

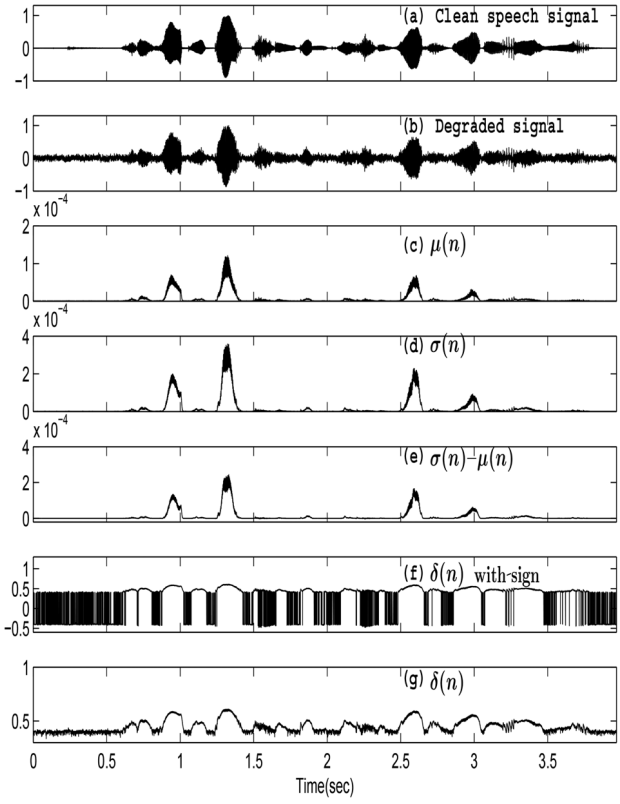


Fig. 3. (a) Clean speech signal. (b) Speech signal corrupted by pink noise at **5 dB** SNR. (c) $\mu(n)$. (d) $\sigma(n)$. (e) $\sigma(n) - \mu(n)$. (f) $\delta(n)$ along with sign. (g) $\delta(n)$.

computation of w_k, the values in the appended silence regions are not considered.

At each time instant, the mean ($\mu(n)$) of the square of the weighted component envelopes computed across frequency corresponds approximately to the energy of the signal at the instant (Fig. 2(c)). The $\mu(n)$ is expected to be higher for speech than for noise in the regions where speech signal is present, as the noise components are deweighted. At each time instant, the standard deviation ($\sigma(n)$) of the square of the weighted component envelopes computed across frequency will also be relatively higher for speech than for noise in the regions of speech due to formant structure (Fig. 2(d)). Hence ($\sigma(n) + \mu(n)$) is generally higher in the speech regions, and lower in the nonspeech regions. Since the spread of noise (after compensation) is expected to be lower, it is observed that the values of ($\sigma(n) - \mu(n)$) are usually lower in the nonspeech regions compared to the values in the speech regions (Fig. 2(e)). Multiplying ($\sigma(n) + \mu(n)$) with ($\sigma(n) - \mu(n)$) gives ($\sigma^2(n) - \mu^2(n)$), which highlights the contrast between speech and nonspeech regions. Figs. 2 and 3 illustrate the features of $\mu(n)$, $\sigma(n)$ and ($\sigma(n) - \mu(n)$) for an utterance corrupted by pink noise at SNR $= -10$ dB and SNR $= 5$ dB, respectively.

Due to large dynamic range of the values of ($\sigma^2(n) - \mu^2(n)$), it is difficult to observe the speech regions with small values of ($\sigma^2(n) - \mu^2(n)$). To highlight the contrast between speech and nonspeech regions, the dynamic range is reduced by computing

$$\delta(n) = \sqrt[M]{|(\sigma^2(n) - \mu^2(n))|}, \tag{13}$$

where M is chosen as 64.

The value of M is not critical. Any value of M in the range of 32 to 256 seems to provide good contrast between speech and nonspeech regions in the plot of $\delta(n)$. In computing $\delta(n)$, only the magnitude of ($\sigma^2(n) - \mu^2(n)$) is considered. If the sign of ($\sigma^2(n) - \mu^2(n)$) is assigned to $\delta(n)$, the values will be fluctuating around zero in the nonspeech regions for most types of noise (see Fig. 2(f) for pink noise), but the short time (20 $-$40 msec) temporal average value will be small and fluctuating, making the noise floor uneven. This makes it difficult to set a threshold for deciding nonspeech regions. The values of $\delta(n)$ will have a high temporal mean value in the nonspeech regions, with small temporal variance (Fig. 2(g)). This helps to set a suitable threshold to isolate nonspeech regions from speech regions. The range of $\delta(n)$ with sign value (Fig. 2(f)) is different from $\delta(n)$ values (Fig. 2(g)). The small temporal spread of $\delta(n)$ values in the nonspeech regions and its mean value helps to fix a suitable threshold. The $\delta(n)$ values in the nonspeech regions is dictated by the noise level. The $\delta(n)$ values in nonspeech regions are high for pink noise degradation at -10 dB SNR (Fig. 2(g)) than at 5 dB SNR (Fig. 3(g)). Note that, by considering the $\delta(n)$ values without sign, we are losing some advantage in the discrimination of nonspeech regions, which has both positive and negative values, compared to speech regions which have mostly positive values. The $\delta(n)$ values with $M = 64$ are used for further processing for decision making. Note the changes in the vertical scales in Figs. 2(f) and 2(g), and also in Figs. 3(f) and 3(g), to understand the significance of using the absolute value, i.e., $\delta(n)$ without sign.

C. Decision Logic

The decision logic is based on $\delta(n)$ for each utterance, by first deriving the threshold over the assumed (20% of the low energy) regions of noise, and then applying the threshold on temporally smoothed $\delta(n)$ values. The window size l_w used for smoothing $\delta(n)$ is adapted based on an estimate of the dynamic range (ρ) of the energy of the noisy signal in each utterance, assuming that there is at least 20% silence region in the utterance. The binary decision of speech and nonspeech at each time instant, denoted as 1 and 0, respectively, is further smoothed (similar to hangover scheme) using an adaptive window, to arrive at the final decision. The following 5 steps describe the implementation details of the decision logic:

1) Computation of threshold (θ):
   Compute the mean (μ_θ) and variance (σ_θ) of the lower 20% of the values of $\delta(n)$ over an utterance. A threshold of $\theta = \mu_\theta + 3\sigma_\theta$ is used in all cases. The θ value depends on each utterance. Thus the threshold value, corresponding to the floor value of $\delta(n)$, is adapted to each utterance, depending on the characteristics of speech and noise in that utterance.
2) Determination of smoothing window l_w:
   The energy E_m of the signal $x(n)$ is computed over a frame of 300 msec for a frame shift of 10 msec, where m is the frame index. The dynamic range (ρ) of the signal is computed as

$$\rho = 10 \log_{10} \frac{\max_m(E_m)}{\min_m(E_m)}. \tag{14}$$

   The window length parameter l_w for smoothing is obtained from the dynamic range (ρ) of the signal. Table III gives the ρ values for degraded speech at SNRs of -10 dB and 5 dB for different noises. The ρ values are high at 5 dB SNR compared to the values at -10 dB SNR for the same noise. The ρ values vary for different noises for the same SNR, because the degradation characteristics of noises vary. For distance speech, the histogram of ρ values for utterances in the C3 case is shown in Fig. 4. The SNR for distant speech depends on the environmental conditions and on the distance of the speaker from microphone. It is observed that the ρ values for the distant speech are spread out, compared to the ρ values for different noises. This is mainly due to the effects of reverberation. The distribution of ρ values depends on the distance as well. The ρ value for each utterance is used to determine some parameter values for further processing of $\delta(n)$ and for arriving at the decision logic. In cases where the $\delta(n)$ represent the discriminating characteristics of speech and nonspeech well, the corresponding ρ values are high, as observed for volvo, leopard and machine gun noises. In such cases, small value of the smoothing window parameter l_w is used. The following values of l_w are chosen based on experimentation with speech degraded by different types of noises at different SNR levels:

$$l_w = 400 \text{ msec, for } \rho < 30. \tag{15}$$
$$l_w = 300 \text{ msec, for } 30 <= \rho <= 40. \tag{16}$$

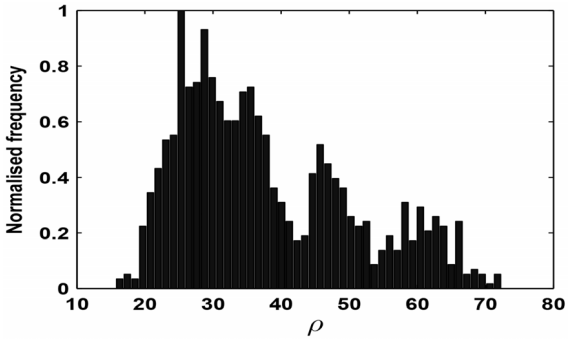


Fig. 4. Histogram of ρ values for distant speech (C3).

TABLE III
VALUES OF ρ FOR SPEECH SIGNAL DEGRADED AT SNRS OF -10 dB AND 5 dB FOR DIFFERENT TYPES OF NOISES. THE VALUE FOR CLEAN SPEECH IS 65.28

NOISE	-10 dB SNR	5 dB SNR
white	14.90	22.61
babble	19.64	36.36
volvo	41.62	56.79
leopard	27.77	43.22
buccaneer1	16.13	28.36
buccaneer2	15.67	22.75
pink	16.73	27.60
hfchannel	16.68	28.46
m109	22.44	35.87
f16	17.84	28.88
factory1	20.48	36.28
factory2	24.13	36.30
machine gun	40.52	64.84

$$l_w = 200 \text{ msec, for } \rho > 40. \tag{17}$$

3) Decision logic at each sampling instant:
   The values of $\delta(n)$ are averaged over a window of size l_w to obtain the averaged value $\bar{\delta}(n)$ at each sample index n. The decision $d(n)$ is made as follows:

$$d(n) = 1, \text{ for } \bar{\delta}(n) > \theta. \tag{18}$$
$$d(n) = 0, \text{ for } \bar{\delta}(n) \leq \theta. \tag{19}$$

4) Smoothing decision at sample level:
   The decision $d(n)$ at each sample is processed over windows of size 300 msec, 400 msec and 600 msec, respectively, for the 3 ranges of ρ indicated in (15), (16) and (17). Let η be the threshold on the proportion (in percentage value) of $d(n)$ values which are 1 in the window. If the percentage of $d(n)$ values which are 1 in the window is above the η value, then the final decision $d_f(n)$ is made 1 at the sampling instant n, otherwise it is 0. The value assigned to η is 60%.
5) Decision at frame level:
   The decision of the AMR methods is given for every 10 msec frame [28]. In order to compare the proposed method with the AMR method, the decision $d_f(n)$ is converted to a 10 msec frame based decision. For each 10 msec nonoverlapping frame, if majority of $d_f(n)$ values are 1, then the frame is marked as speech, otherwise it is marked as nonspeech. The ground truth of speech signals is also derived for each 10 msec frame.

TABLE IV
AVERAGED SCORES ACROSS ALL NOISE TYPES FOR TWO SNR LEVELS FOR TIMIT DATABASE

SNR (dB)	Method	CORRECT	FEC	MSC	OVER	NDS
-10	Proposed	**79.11**	0.06	15.21	0.03	5.49
	AMR2	72.60	0.07	19.31	0.04	7.86
	AMR1	51.70	0.02	2.50	0.10	45.56
5	Proposed	**95.36**	0.02	2.27	0.05	2.20
	AMR2	88.85	0.04	2.31	0.09	8.58
	AMR1	76.05	0.04	1.52	0.09	22.17

V. EVALUATION OF PROPOSED APPROACH

The proposed method is compared with the state-of-the-art AMR1 and AMR2 methods [28]. AMR1 and AMR2 methods extract subband energies using filter banks. Several acoustical features like pitch, tone, etc., are used to arrive at the decision. Post-processing techniques like hangover are also used [22]. In this paper, we use the version 3GPP TS 26.104 of the AMR methods [28].

We use 5 parameters to evaluate our approach against AMR methods [29] for comparison.

- CORRECT: Correct decisions made by the VAD.
- FEC (front end clipping): Clipping due to speech being misclassified as noise in passing from noise to speech activity.
- MSC (mid speech clipping): Clipping due to speech being misclassified as noise during a speech region.
- OVER (carry over): Noise interpreted as speech in passing from speech activity to noise.
- NDS (noise detected as speech): Noise interpreted as speech within silence/noise region.

All the above parameters are divided by the total number of frames (both speech and nonspeech frames), and then multiplied by 100 to get the percentage value (%). Combining FEC and MSC gives true rejection (TR). Combining OVER and NDS gives false acceptance (FA). The TR indicates the percentage of speech regions not detected as speech, whereas the FA indicates the percentage of nonspeech regions accepted as speech. For good performance, CORRECT should be high, and both TR and FA should be low.

The AMR2 method performs better than AMR1 method in all cases, which is evident from the averaged scores across all noise types for the two different SNRs given in Table IV. Hence we only consider AMR2 scores for comparison.

Tables V, VI, VII and VIII show the performance of the proposed method in comparison with the AMR2 method for different type of degradations and at different SNR conditions. The best performance in each case is indicated by boldface for CORRECT score.

In the following, the performance of the proposed method is discussed for different types of degradation.

A. Performance on TIMIT Database for Different Types of Noises

Performance of the proposed method under different noise conditions of NOISEX database is given in Table V for two different SNR values, i.e., at -10 dB and 5 dB.

TABLE V
RESULTS FOR TIMIT DATABASE FOR DIFFERENT TYPES OF NOISES AT TWO SNR LEVELS IN COMPARISON WITH AMR2 METHOD

NOISE (SNR in dB)	Method	CORRECT	FEC	MSC	OVER	NDS
white (-10)	Proposed	**77.60**	0.08	21.99	0.01	0.23
	AMR2	63.23	0.11	34.32	0.01	2.24
white (5)	Proposed	**97.01**	0.02	1.81	0.05	1.04
	AMR2	87.47	0.08	8.55	0.06	3.74
babble (-10)	Proposed	**67.72**	0.04	12.06	0.05	20.04
	AMR2	61.67	0.05	13.10	0.07	25.01
babble (5)	Proposed	**93.27**	0.03	2.56	0.05	4.01
	AMR2	72.43	0.03	0.52	0.11	26.80
volvo (-10)	Proposed	**98.04**	0.02	0.53	0.08	1.26
	AMR2	95.93	0.02	0.24	0.11	3.59
volvo (5)	Proposed	**96.39**	0.04	2.38	0.06	1.03
	AMR2	94.37	0.00	0.54	0.11	4.89
leopard (-10)	Proposed	**97.09**	0.02	1.18	0.06	1.58
	AMR2	95.92	0.05	0.88	0.10	2.95
leopard (5)	Proposed	**97.82**	0.02	0.78	0.07	1.22
	AMR2	95.61	0.01	0.16	0.11	4.00
buccaneer1 (-10)	Proposed	**69.76**	0.09	25.90	0.01	4.15
	AMR2	65.97	0.11	33.10	0.01	0.71
buccaneer1 (5)	Proposed	**95.59**	0.02	2.23	0.05	2.03
	AMR2	93.92	0.07	3.81	0.08	2.01
buccaneer2 (-10)	Proposed	**76.54**	0.08	21.41	0.01	1.87
	AMR2	64.46	0.11	34.00	0.00	1.33
buccaneer2 (5)	Proposed	**96.89**	0.02	1.81	0.05	1.16
	AMR2	90.78	0.07	6.28	0.06	2.69
pink (-10)	Proposed	**74.23**	0.09	25.43	0.01	0.16
	AMR2	66.79	0.11	32.30	0.00	0.69
pink (5)	Proposed	**97.07**	0.02	1.75	0.05	1.04
	AMR2	94.52	0.07	3.22	0.08	1.99
hfchannel (-10)	Proposed	**75.03**	0.08	23.48	0.02	1.30
	AMR2	71.22	0.09	27.22	0.03	1.33
hfchannel (5)	Proposed	**96.94**	0.02	1.57	0.06	1.35
	AMR2	94.69	0.05	2.57	0.09	2.49
m109 (-10)	Proposed	**89.68**	0.04	6.49	0.04	3.69
	AMR2	82.80	0.08	15.03	0.04	1.94
m109 (5)	Proposed	**97.32**	0.01	0.72	0.07	1.81
	AMR2	95.63	0.04	0.40	0.11	3.70
f16 (-10)	Proposed	**75.94**	0.08	22.37	0.02	1.51
	AMR2	69.88	0.10	29.18	0.01	0.72
f16 (5)	Proposed	**97.10**	0.02	1.43	0.06	1.34
	AMR2	95.64	0.06	2.06	0.09	2.04
factory1 (-10)	Proposed	**67.56**	0.05	13.53	0.05	18.74
	AMR2	58.80	0.06	17.42	0.06	23.57
factory1 (5)	Proposed	**91.72**	0.02	1.85	0.06	6.28
	AMR2	74.12	0.04	1.42	0.10	24.21
factory2 (-10)	Proposed	**82.20**	0.05	9.55	0.04	8.09
	AMR2	82.16	0.07	14.02	0.06	3.59
factory2 (5)	Proposed	**95.09**	0.02	0.91	0.07	3.85
	AMR2	94.45	0.04	0.36	0.11	4.94
machine gun (-10)	Proposed	**77.13**	0.07	13.85	0.03	8.81
	AMR2	64.97	0.01	0.26	0.11	34.56
machine gun (5)	Proposed	**87.55**	0.08	9.78	0.03	2.45
	AMR2	71.43	0.00	0.24	0.11	28.12

TABLE VI
RESULTS FOR NTIMIT AND CTIMIT DATABASE IN COMPARISON WITH AMR2 METHOD

Data	Method	CORRECT	FEC	MSC	OVER	NDS
NTIMIT	Proposed	**94.78**	0.02	1.66	0.18	3.18
	AMR2	93.59	0.12	2.91	0.21	2.91
CTIMIT	Proposed	**91.14**	0.04	4.21	0.15	4.31
	AMR2	87.68	0.15	8.28	0.19	3.46

TABLE VII
RESULTS FOR DISTANT SPEECH FOR DIFFERENT VALUES OF η IN THE DECISION LOGIC IN COMPARISON WITH AMR2 METHOD

Data	Method	CORRECT	FEC	MSC	OVER	NDS
C0	η=60%	84.54	0.00	0.11	0.26	14.92
	η=70%	86.73	0.02	0.26	0.23	12.61
	η=80%	88.34	0.07	0.67	0.20	10.57
	η=90%	**88.93**	0.10	1.64	0.17	8.99
	AMR2	88.87	0.07	0.39	0.26	10.23
C1	η=60%	89.40	0.01	0.56	0.22	9.65
	η=70%	90.91	0.04	1.20	0.17	7.53
	η=80%	91.03	0.13	2.53	0.14	6.02
	η=90%	89.76	0.14	4.53	0.12	5.27
	AMR2	**91.40**	0.11	0.67	0.25	7.35
C2	η=60%	87.03	0.02	0.87	0.22	11.71
	η=70%	88.50	0.06	1.70	0.18	9.42
	η=80%	**88.63**	0.13	3.14	0.15	7.78
	η=90%	87.91	0.14	5.21	0.13	6.44
	AMR2	87.81	0.11	0.98	0.25	10.64
C3	η=60%	86.89	0.03	1.56	0.21	11.16
	η=70%	**87.89**	0.07	2.77	0.17	8.95
	η=80%	87.58	0.13	4.61	0.14	7.37
	η=90%	86.12	0.14	7.35	0.12	6.08
	AMR2	87.61	0.13	2.99	0.23	8.82

TABLE VIII
RESULTS FOR TIMIT CLEAN CASE FOR DIFFERENT VALUES OF η IN THE DECISION LOGIC IN COMPARISON WITH AMR2 METHOD

Method	CORRECT	FEC	MSC	OVER	NDS
η=40%	**95.72**	0.02	2.71	0.06	1.37
η=50%	94.92	0.05	3.96	0.06	0.92
η=60%	93.55	0.07	5.61	0.04	0.62
η=70%	91.66	0.08	7.66	0.04	0.46
η=80%	89.42	0.09	10.01	0.03	0.35
AMR2	93.59	0.12	2.91	0.21	2.91

It is observed that performance of the proposed method is higher than that of AMR2 method for all types of noises. For five types of noises, the performance is illustrated for an utterance in the form of plots shown in Figs. 5 and 6 at SNRs of -10 dB and 5 dB, respectively. For each type of noise, the degraded signal, the corresponding $\delta(n)$ values and the derived VAD decision (thick line) are shown. In addition, the AMR2 decision is also shown by thin solid lines for comparison. The ground truth is marked in Figs. 5(a) and 6(a) by a thin line.

As can be seen from Figs. 5(a) and 6(a) for white noise case, many speech regions are missed in the AMR2 method, resulting in high TR. In the case of babble noise at 5 dB SNR, the features in the speech regions stand out over the nonspeech regions (Fig. 6(e)), and hence the FA is lower for the proposed method than for the AMR2 method (Fig. 6(f)).

Since most of the energy is concentrated in the low frequency regions for volvo noise, it is relatively easier to reduce the effect of this type of noise, and hence the proposed method performs better at the two noise levels (Figs. 5(i) and 6(i)).

A significant lower TR is seen in the case of pink noise for the proposed method compared to the AMR2 method. This is due to attenuation of noise regions by weighting (Figs. 5(l)). This can also be seen in the 3D plots given in Fig. 1.

Due to its high temporal variance, most VAD algorithms detect the machine gun chunks as speech. The high temporal resolution of the features in the proposed method gives better performance for the proposed method than for the AMR2 method as indicated in Table V. It is interesting to see in Figs. 5(o) and 6(o) that the nonspeech regions affected by the machine gun noise are identified as nonspeech by the proposed

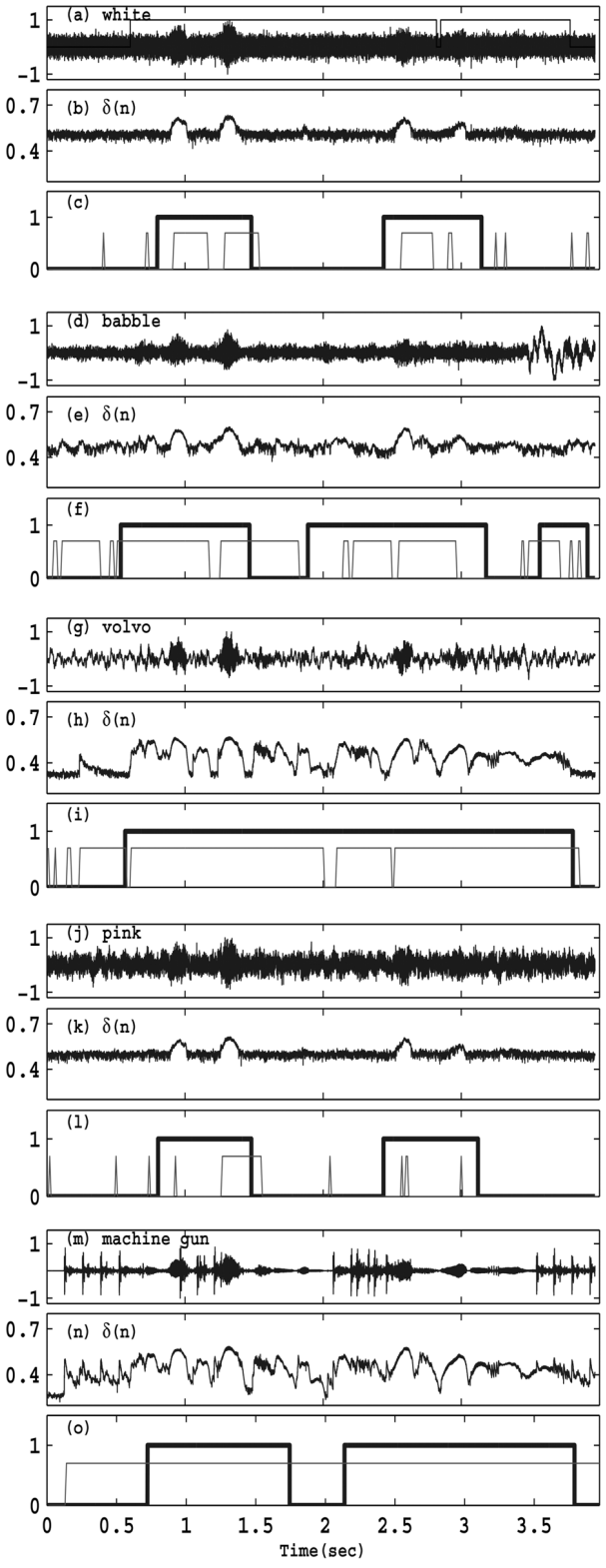


Fig. 5. Illustration of results of VAD for different types of NOISEX data at -10 dB SNR. Each noise type has three subfigures: Degraded signal at -10 dB SNR, $\delta(n)$, and decision for the proposed method (thick line) and for AMR2 method (thin line). White noise (a, b, c), Babble noise (d, e, f), Volvo noise (g, h, i), Pink noise (j, k, l), Machine gun noise (m, n, o). The ground truth is indicated on top of the degraded speech signal in (a).

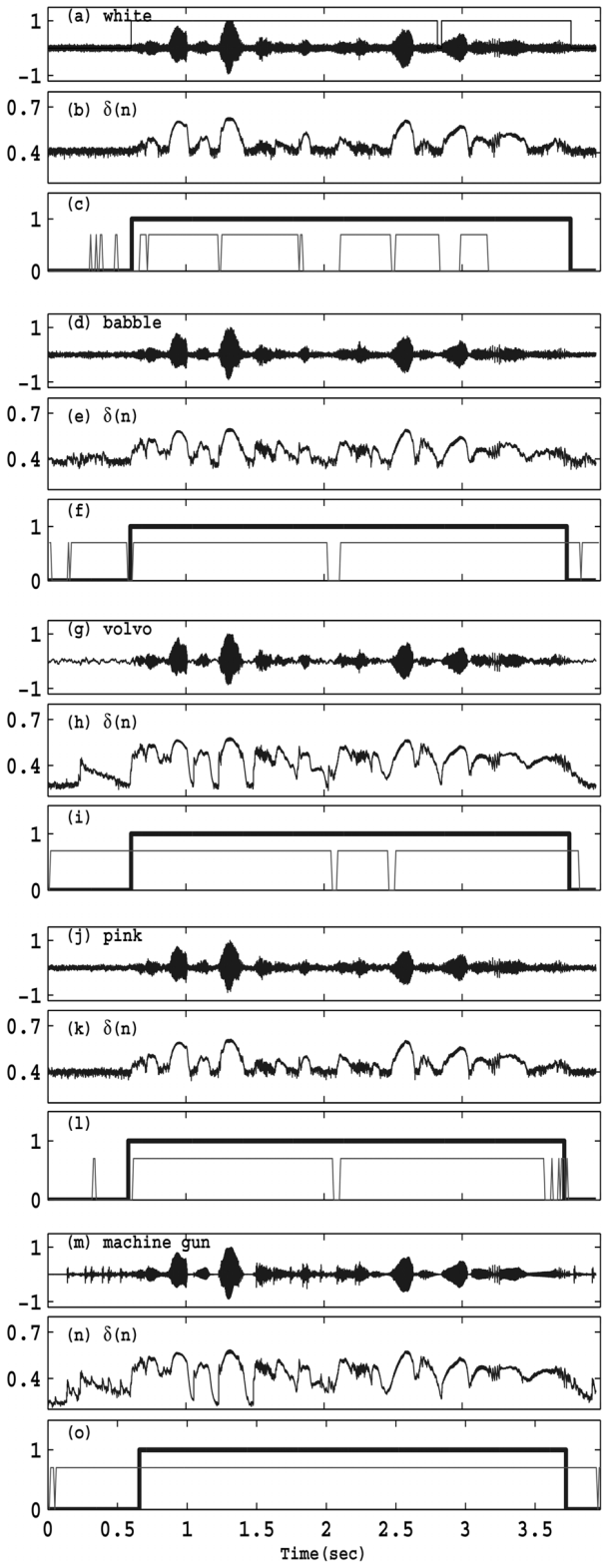


Fig. 6. Illustration of results of VAD for different types of NOISEX data at 5 dB SNR. Each noise type has three subfigures: Degraded signal at 5 dB SNR, $\delta(n)$, and decision for the proposed method (thick line) and for AMR2 method (thin line). White noise (a, b, c), Babble noise (d, e, f), Volvo noise (g, h, i), Pink noise (j, k, l), Machine gun noise (m, n, o). The ground truth is indicated on top of the degraded speech signal in (a).

method, whereas the AMR2 method accepts them as speech. The LTSV method proposed in [8] shows poor performance

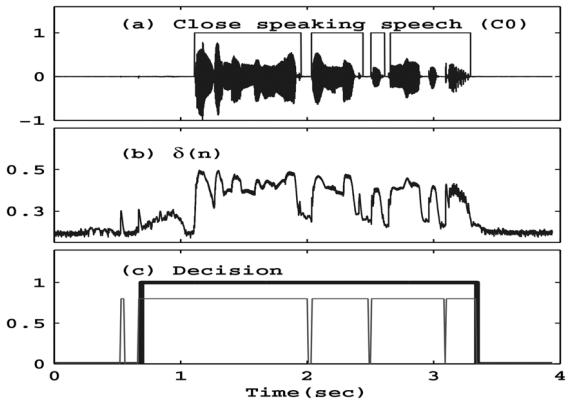


Fig. 7. (a) Close speaking speech (C0) with ground truth indicated on top. (b) $\delta(n)$. (c) Decision of the proposed method at $\eta = 90\%$ (thick line) and the AMR2 method (thin line).

for this noise. The multi-band LTSV method [9] also fails to discriminate transient noise from speech.

B. Performance on NTIMIT and CTIMIT Databases

Performance of the proposed method is similar to the AMR2 method for the NTIMIT data (Table VI), and is higher than for the CTIMIT data (Table VI). This may be due to the cellphone (coding) effects, which degrade speech more than the telephone channel (NTIMIT). The l_w value is 200 msec for most of the utterances in these cases because of high ρ value (see (17)).

C. Performance on Distant Speech

Distant speech is an amalgam of unknown degradations, and the data for a given environment may be limited. The reverberation present in the distant speech signals has high variance in the time domain, as does the speech. So VAD algorithms often confuse reverberation component for speech. The VAD algorithms which bank on temporal variance ([8]) may not perform well, because the distant speech is highly nonstationary, and even the nonspeech regions may have significant temporal variance.

Fig. 7 illustrates the decision obtained by the proposed method and by the AMR2 method for the case of close speaking speech (C0). The errors in the AMR2 method and the proposed method are mostly due to FA (Fig. 7(c)). Note that the $\delta(n)$ values (Fig. 7(b)) have large fluctuations in the speech region, and also it has low floor values as for any clean speech. It is to be noted, that for distant microphone case the performance of the proposed method gives results similar to the AMR2 method, indicating that the proposed method does not fail. Table VII indicates that by proper choice of the value of the η parameter, there can be slight improvement. But the improvement may not be significant. The interesting aspect is that most of the errors in this case are due to false acceptance (FA). This occurs because the degradation in silence regions is not uniform in the case of distant speech, making it difficult to set proper threshold either in the proposed method or in the AMR2 method. One would notice larger fluctuations in the values of $\delta(n)$ in the nonspeech regions, which would result in higher FA rate. It appears that reverberant effects also may be playing a significant role in producing large fluctuations in the values of $\delta(n)$, as it is difficult to compensate those effects by noise deweighting.

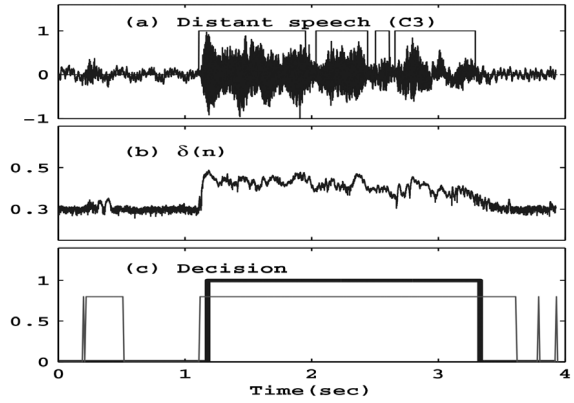


Fig. 8. (a) Distance speech (C3) with ground truth indicated on top. (b) $\delta(n)$. (c) Decision of the proposed method at $\eta = 90\%$ (thick line) and the AMR2 method (thin line).

It is also interesting to note that even for the relatively cleaner speech (i.e., C0 case in distant speech), there will be large fluctuations in the $\delta(n)$ values in the silence regions, making it difficult to set the thresholds properly. Hence the performance by both the proposed method and the AMR2 method is poorer for C0 case than for the more degraded case of C1.

Fig. 8 illustrates the decision obtained by the proposed method and by the AMR2 method for the distance speech (case C3) for the same utterance shown in Fig. 7. The error is mostly in FA for the AMR2 method (Fig. 8(c)). Note that the $\delta(n)$ values (Fig. 8(b)) have lower dynamic range in the speech region. Also, it has high floor value, as for most degraded speech.

Performance of distant speech can be improved by increasing the η value, as it reduces FA. Table VII shows the improvement in the performance of the distant speech with increase in the η value for the proposed method in comparison with the AMR2 method. Note that the large values of η can also cause increase in the true rejection (TR), which may result in overall reduction in correct decision.

D. Performance on TIMIT Database for Clean Speech

Performance of the proposed method on clean speech is given in Table VIII. It is interesting to note that smoothing and threshold logic for degraded speech smear the information across time, thus reducing the temporal resolution of the final decision. Hence when the decision logic is applied to clean data, it appears to give poor performance. Due to the high dynamic range in both time and frequency domains, the clean speech signal needs to be treated differently in order to obtain good performance.

In contrast to the C0 case of distant speech, for the clean TIMIT data, the error is more in the true rejection (TR) as in Table VIII. This is because for the clean TIMIT data in the silence region, the $\delta(n)$ values are very low and are fluctuating, making it difficult to set the proper threshold. In this case the TR can be reduced by reducing the threshold value, or equivalently reducing the η value.

The scores given in Tables V and VI are for fixed values of the parameters in the decision logic (Section IV-C). The η value has been fixed at 60% for most of the cases. A better performance may be achieved, if the parameters θ, η, l_w are adapted suitably for each type of degradation.

TABLE IX
AVERAGED SCORES ACROSS DIFFERENT NOISE TYPES FOR TWO SNR LEVELS FOR TIMIT DATABASE

SNR (dB)	Method	CORRECT	FEC	MSC	OVER	NDS
-[0	Proposed	81.25	0.05	12.36	0.03	6.20
	DFT	**81.69**	0.02	4.31	0.06	13.83
	Gammatone	81.63	0.06	11.98	0.03	6.20
	AMR2	74.63	0.06	15.19	0.05	9.94
5	Proposed	**94.77**	0.02	2.37	0.05	2.68
	DFT	93.38	0.01	1.12	0.06	5.33
	Gammatone	93.24	0.03	3.78	0.03	2.82
	AMR2	88.75	0.04	1.52	0.09	9.46

TABLE X
AVERAGED SCORES ACROSS ALL NOISE TYPES FOR TWO SNR LEVELS OF UNWEIGHTED AND WEIGHTED SFF OUTPUT FOR TIMIT DATABASE

SNR (dB)	Method	CORRECT	FEC	MSC	OVER	NDS
-10	Unweighted SFF	70.73	0.07	19.45	0.03	9.60
	Weighted SFF	**78.04**	0.07	15.82	0.03	5.94
5	Unweighted SFF	93.18	0.03	3.80	0.05	2.83
	Weighted SFF	**95.15**	0.02	2.25	0.06	2.41

E. Performance Comparison with DFT and Gammatone Filters

The proposed method is evaluated using filterbank energy contours using DFT and 128 gammatone filters [21]. After deriving the band energy contours, the subsequent processing, including weighting, the energy contours, computation of $\delta(n)$, thresholding and decision logic, are all same in these cases as in the SFF method described before.

The results are given in Table IX in terms of averaged performance over 11 different noise types (except white and pink noises), using 50 utterances of TIMIT data, for two different noise levels (-10 dB and 5 dB). It is interesting to note that all the three methods of preprocessing namely, SFF, DFT and gammatone filters, give similar results. All of them are significantly better than the results using the AMR2 method.

Note that the three methods of preprocessing may perform differently for different noise types. We have observed that for synthetic noises like white and pink noises, the performance by DFT and gammatone filtering is better than by SFF. This is due to some temporal and spectral averaging of noises in the high frequency region ($>$ 2000 Hz) due to temporal averaging in the case of DFT and due to spectral smoothing in the case of gammatone filters. The performance improvement for all the three methods will be similar even for these two types of noises, if in the SFF method some smoothing is done in the time and frequency domains, especially in the higher frequency region, before computing mean and variance across frequency. Note that the performance improvement of these three preprocessing methods over the AMR2 method is due to the subsequent processing of the energy contours in each band, especially the weighting in (12). The effect of weighting can be seen in the performance of the proposed method with and without weighting as given in Table X. The average scores across all noise types for two different SNR values (-10 dB and 5 dB) are given using unweighted and weighted SFF output for 50 utterances of TIMIT data.

VI. Summary

A new VAD method is proposed based on single frequency filtering (SFF) approach introduced in this paper. The method exploits the fact that speech has high SNR regions at different frequencies and at different times. The variance of speech across frequency is higher than that for noise, after compensating for spectral characteristics for noise. The spectral characteristics of noise are determined using the floor of the temporal envelope at each frequency, computed by the SFF approach.

The $\delta(n)$ feature proposed for VAD decision is robust against degradation, as evidenced by the high CORRECT percentage scores obtained for all types of noises. The proposed method is tested over standard TIMIT, NTIMIT and CTIMIT databases, as well as for distance speech, thus covering varieties of degradations.

While the results show significant improvement in performance of the proposed method, in comparison with the AMR2 method, better results may be obtained, if the decision logic parameters (θ, η, l_w) are made degradation-specific. It was noticed that adapting the parameters θ, η, l_w based on the degradation characteristics estimated from ρ has improved the overall performance. Adapting the threshold with time in each utterance may also improve the performance. Further improvement can be expected if other characteristics of speech, such as voicing, are also included in the decision logic.

The SFF method yields envelopes at any desired frequency, with high temporal and spectral resolution. This property can be exploited for many other applications in speech processing, such as robust pitch extraction, speech enhancement, and deriving robust features for speech and speaker recognition. Our preliminary studies indicate that the SFF method is indeed showing promise in some of these applications.

Acknowledgment

The authors thank the members of ECESS Consortium Siemens AG, Corporate Technology, Germany, for granting permission to use SPEECON database.

References

[1] T. Kinnunen, E. Chernenko, M. Tuononen, P. Fränti, and H. Li, "Voice activity detection using MFCC features and support vector machine," in *Proc. Int. Conf. Speech Comput. (SPECOM'07)*, 2007, vol. 2, pp. 556–561.

[2] I. McCowan, D. Dean, M. McLaren, R. Vogt, and S. Sridharan, "The delta-phase spectrum with application to voice activity detection and speaker recognition," *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 19, no. 7, pp. 2026–2038, Sep. 2011.

[3] J. Haigh and J. Mason, "Robust voice activity detection using cepstral features," in *Proc. IEEE TENCON'93*, 1993, pp. 321–324.

[4] N. Dhananjaya and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," *IEEE Signal Process. Lett.*, vol. 17, no. 3, pp. 273–276, Mar. 2010.

[5] T. Pham, M. Stark, and E. Rank, "Performance analysis of wavelet subband based voice activity detection in cocktail party environment," in *Proc. Int. Conf. Comput. Commun. Technol.*, Oct. 2010, pp. 85–88.

[6] Z. Song, T. Zhang, D. Zhang, and T. Song, "Voice activity detection using higher-order statistics in the Teager energy domain," in *Proc. Wireless Commun. Signal Process.*, Nov. 2009, pp. 1–5.

[7] J. Ramrez, J. C. Segura, C. Bentez, A. D. L. Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Commun.*, vol. 42, pp. 3–4, 2004.

[8] P. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 19, no. 3, pp. 600–613, Mar. 2011.

[9] A. Tsiartas, T. Chaspari, N. Katsamanis, P. K. Ghosh, M. Li, M. Van Segbroeck, A. Potamianos, and S. Narayanan, "Multi-band long-term signal variability features for robust voice activity detection," in *Proc. Interspeech*, Aug. 2013, pp. 718–722.

[10] M. Van Segbroeck, A. Tsiartas, and S. Narayanan, "A robust frontend for VAD: Exploiting contextual, discriminative and spectral cues of human voice," in *Proc. Interspeech*, Aug. 2013, pp. 704–708.

[11] Y. W. Jitong Chen and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE Trans. Speech Audio Process.*, vol. 22, no. 12, pp. 1993–2002, Dec. 2014.

[12] X.-L. Zhang and D. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *Proc. Interspeech*, Sep. 2014, pp. 1534–1538.

[13] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 14, no. 2, pp. 412–424, Mar. 2006.

[14] T. Pham, C. Tang, and M. Stadtschnitzer, "Using artificial neural network for robust voice activity detection under adverse conditions," in *Proc. Int. Conf. Comput. Commun. Technol., RIVF*, Jul. 2009, pp. 1–8.

[15] D. Ying, Y. Yan, J. Dang, and F. Soong, "Voice activity detection based on an unsupervised learning framework," *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 19, no. 8, pp. 2624–2633, Nov. 2011.

[16] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 21, no. 4, pp. 697–710, 2013.

[17] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[18] J. Ramirez, J. Segura, C. Benitez, A. De la Torre, and A. Rubio, "An effective subband OSF-based VAD with noise reduction for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 6, pp. 1119–1129, Nov. 2005.

[19] P. Harding and B. Milner, "On the use of machine learning methods for speech and voicing classification," in *Proc. Interspeech*, Sep. 2012.

[20] F. Germain, D. L. Sun, and G. J. Mysore, "Speaker and noise independent voice activity detection," in *Proc. Interspeech*, Aug 2013, pp. 732–736.

[21] V. Hohmann, "Frequency analysis and synthesis using a gammatone filterbank," *Acta Acust. United with Acust.* , vol. 88, no. 3, pp. 433–442, Jan. 2002.

[22] ETSI, Voice activity detector (VAD) for adaptive multirate (AMR) speech traffic channels, ETSI EN 301 708 v.7.1.1, Dec. 1999.

[23] J. S. Garofolo *et al., TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia, PA, USA: Linguistic Data Consortium, 1993.

[24] A. Varga and J. H. Steeneken, "Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[25] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," in *Proc. ICASSP*, Apr. 1990, pp. 109–112.

[26] K. Brown and E. George, "CTIMIT: A speech corpus for the cellular environment with applications to automatic speech recognition," in *Proc. ICASSP*, May 1995, pp. 105–108.

[27] R. Siemund, H. Huge, S. Kunzmann, and K. Marasek, "SPEECON - speech data for consumer devices," in *Proc. LREC*, 2000, pp. 883–886.

[28] [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/26104.htm

[29] D. Freeman, G. Cosier, C. Southcott, and I. Boyd, "The voice activity detector for the Pan-European digital cellular mobile telephone service," in *Proc. ICASSP*, May 1989, pp. 369–372.

G. Aneeja received the B.Tech. in electronics and communications engineering from Jawaharlal Nehru Technological University, Hyderabad, India, in 2009. She is currently pursuing the Ph.D. degree at the International Institute of Information Technology, Hyderabad, India. Her research interests include signal processing, speech analysis, voice activity detection and speaker verification.

B. Yegnanarayana (M'78–SM'84–F'13) received the B.Sc. degree from Andhra University, Waltair, India, in 1961, and the B.E., M.E., and Ph.D. degrees in electrical communication engineering from the Indian Institute of Science (IISc) Bangalore, India, in 1964, 1966, and 1974, respectively. He is currently an Institute Professor at the International Institute of Information Technology (IIIT), Hyderabad. He was Professor and Microsoft chair at IIIT, Hyderabad, from 2006 to 2012. Prior to joining IIIT, Hyderabad, he was a Professor at the Indian Institute of Technology, Madras, India (1980 to 2006), a Visiting Associate Professor at Carnegie-Mellon University, Pittsburgh, USA (1977 to 1980), and a member of the faculty at the IISc, Bangalore (1966 to 1978). His research interests are in signal processing, speech, image processing, and neural networks. He has published over 350 papers in these areas in IEEE journals and other international journals, and in the proceedings of national and international conferences. He is also the author of the book *Artificial Neural Networks* (Prentice-Hall of India, 1999). He has supervised 30 Ph.D. dissertations and 36 M.S. theses. He is a Fellow of the Indian National Academy of Engineering (INAE), a Fellow of the Indian National Science Academy (INSA), a Fellow of the Indian Academy of Sciences, a Fellow of IEEE (USA) and a Fellow of the International Speech Communications Association (ISCA). He was the recipient of the third IETE Prof. S. V. C. Aiya Memorial Award in 1996. He received the Prof. S. N. Mitra memorial Award for the year 2006 from the INAE. He was awarded the 2013 Distinguished Alumnus award from IISc, Bangalore. He was awarded the "Sayed Husain Zaheer Medal (2014)" of INSA in 2014. He was the Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING during 2003–2006. Currently he is an Associate Editor for Springer's international journal *Circuits, Systems and Signal Processing*.

Study of the effects of vocal tract constriction on glottal vibration

Vinay Kumar Mittal,[a] B. Yegnanarayana, and Peri Bhaskararao
Speech and Vision Laboratory, International Institute of Information Technology, Hyderabad - 500032, India

(Received 1 November 2013; revised 26 July 2014; accepted 15 August 2014)

Characteristics of glottal vibration are affected by the obstruction to the flow of air through the vocal tract system. The obstruction to the airflow is determined by the nature, location, and extent of constriction in the vocal tract during production of voiced sounds. The effects of constriction on glottal vibration are examined for six different categories of speech sounds having varying degree of constriction. The effects are examined in terms of source and system features derived from the speech and electroglottograph signals. It is observed that a high degree of constriction causing obstruction to the flow of air results in large changes in these features, relative to the adjacent steady vowel regions, as in the case of apical trill and alveolar fricative sounds. These changes are insignificant when the obstruction to the airflow is less, as in the case of velar fricative and lateral approximant sounds. There are no changes in the excitation features when there is a free flow of air along the auxiliary tract, despite constriction in the vocal tract, as in the case of nasals. These studies show that effects of constriction can indeed be observed in the features of glottal vibration as well as vocal tract resonances. © *2014 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4894789]

PACS number(s): 43.72.Ar [CYE] Pages: 1932–1941

I. INTRODUCTION

Speech is produced by exciting the time-varying vocal tract system. The major source of excitation is the quasi-periodic vibration of the vocal folds at the glottis (Fant, 2004), referred to as *voicing*. The mode of glottal vibration can be controlled voluntarily for producing different phonation types such as modal, breathy, and creaky voices (Laver, 1994; Ladefoged and Johnson, 2011). The rate of glottal vibration can also be controlled *voluntarily*, giving rise to changes in pitch. Glottal vibration may also be affected due to coupling of the vocal tract system with the glottis. These changes in the glottal vibration may be viewed as *involuntary*.

In the production of some speech sounds, the source of excitation is affected due to coupling of the vocal tract system with the glottis. The interaction of glottal source and the vocal tract system has been studied by several researchers (Fant and Lin, 1987; Chi and Sonderegger, 2007; Titze *et al*., 2008; Titze, 2008; Lucero *et al*., 2012). The involuntary changes in the glottal vibration occur during changes in the "intrinsic pitch" (fundamental frequency F_0) of some high vowels (Ewan and Ohala, 1979; Shadle, 1985; Ohala and Eukel, 1987). The effect could be due to either coupling between the vocal tract and glottis (Ewan, 1977; Ohala and Eukel, 1987), or due to tongue-pull effect (Ewan and Ohala, 1979; Ohala and Eukel, 1987). The effects of coupling between oral and subglottal cavities were examined through vowel formants (Perkell and Cohen, 1989; Sonderegger, 2004; Chi and Sonderegger, 2007). Discontinuity in the second formant frequency and signal attenuation were observed in diphthongs near the subglottal resonance in the range of 1280–1620 Hz, due to subglottal coupling (Chi and Sonderegger, 2007). Subglottal resonances were also measured in the case of nasalization (Stevens *et al*., 1975). Studies on the source-system interaction were also carried out for other categories of speech sounds, such as fricatives and stops (Stevens, 1971).

Other studies on source-tract interaction focused mainly on the physical aspects (Stevens, 1977; Titze and Story, 1997; Hatzikirou *et al*., 2006; Zhang *et al*., 2006). The physical models of the acoustic interaction of the voice source with subglottal vocal tract system were studied in Titze (1988), Titze and Story (1997), and Hatzikirou *et al*. (2006). The effect of glottal opening on the response of the vocal tract system was studied in Fant and Lin (1987), Barney *et al*. (1999, 2007), and Ruty *et al*. (2008). The nonlinear phenomenon due to source-tract coupling is related to the air flow across glottis during phonation (Rothenberg, 1981; Chan and Titze, 2006; Zhang *et al*., 2006; Titze, 2008; Lucero *et al*., 2012). The source-tract interaction was observed to induce, under certain circumstances, some complex voice instabilities, such as sudden frequency jumps, subharmonic generation, and random changes in frequency, especially during F_0 and F_1 (i.e., fundamental and first formant frequencies) crossovers (Hatzikirou *et al*., 2006; Titze *et al*., 2008; Titze, 2008).

In the current study, we examine the effect of degree and location of the constriction of the vocal tract system on the glottal vibration for a selected set of six categories of voiced consonant sounds, namely, apical trill, alveolar fricative, velar fricative, apical lateral approximant, alveolar nasal, and velar nasal. The degree of constriction to the air flow is determined by the size, type, and location of the stricture in the vocal tract. These consonant sounds are considered in the context of vowel [a]. Three types of occurrences, namely, single, geminated, and prolonged are examined for each of the six categories of sounds. The speech signal along with the electroglottograph (EGG) signal (Fourcin and Abberton, 1971; Fant *et al*., 1985) is used for analysis of

[a]Author to whom correspondence should be addressed. Electronic mail: vinay.mittal@iiit.ac.in

 0001-4966/2014/136(4)/1932/10/$30.00

these sounds. Changes in the system characteristics are analyzed using two dominant peak frequencies (F_{D_1} and F_{D_2}) derived using linear prediction (LP) analysis (Makhoul, 1975). Source features such as the instantaneous fundamental frequency (F_0) and strength of impulse-like excitation (*SoE*) are extracted from the speech signal using the zero-frequency filtering (ZFF) method (Murty and Yegnanarayana, 2008; Yegnanarayana and Murty, 2009).

The paper is organized as follows. In Sec. II, some production details of the strictures for the six categories of sounds are discussed. In Sec. III, the details of data collection are given. Features extracted from the speech signal are discussed in Sec. IV. In Sec. V, the changes in the glottal source characteristics and associated changes in the vocal tract system characteristics are analyzed qualitatively in terms of the features derived from both the EGG and speech signals, for geminated occurrences of the different categories of sounds. In Sec. VI, the effects of vocal tract constriction on the glottal vibration characteristics are discussed using quantitative changes in the features derived from the speech signals, for all three types of occurrences of the six categories of sounds. In Sec. VII, a summary is given along with a scope for further work in this research area.

II. STRICTURES IN THE VOCAL TRACT

The different sound categories selected for study in this paper differ in cross-sectional area of stricture, besides place of articulation, i.e., location of stricture (Fant, 1970; Catford, 2001) during production. Differences in the stricture for stop, trill, fricative, and approximant sounds are schematically represented in Figs. 1(a)–1(d), respectively (Catford, 2001). In the production of apical trill ([r]) sound, the oral stricture opens and closes periodically [as shown in Fig. 1(b)], at the rate of 25–50 Hz (McGowan, 1992; Dhananjaya *et al.*, 2012). This periodic opening and closing

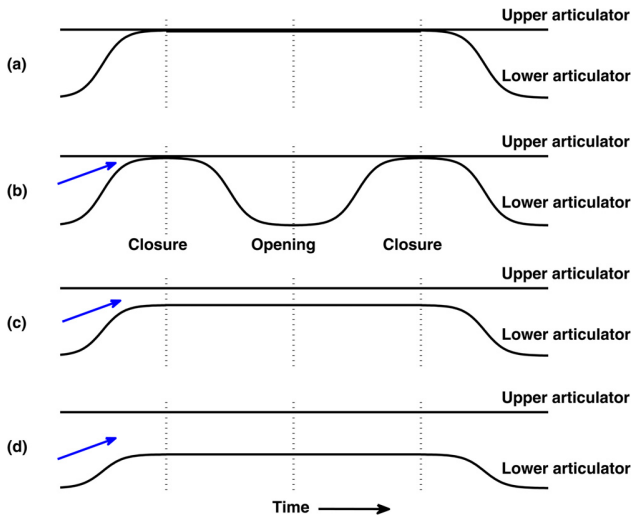


FIG. 1. (Color online) Illustration of strictures for voiced sounds: (a) stop, (b) trill, (c) fricative, and (d) approximant. Relative difference in the stricture size between upper articulator (teeth or alveolar/palatal/velar regions of palate) and lower articulator (different areas of tongue) is shown schematically, for each case. Arrows indicate the direction of air flow passing through the vocal tract.

of the oral cavity produces time-varying constriction in the vocal tract. In recent studies on the production of apical trills, the effect of the system on the source characteristics (Dhananjaya *et al.*, 2012) and the role of source-system coupling (Mittal *et al.*, 2012) were examined. In this paper, we examine the excitation source characteristics of apical trills ([r]) using the EGG signal along with the speech signal.

Production of fricatives involves narrow constriction of the vocal tract at some point along its length [Fig. 1(c)], which may affect the glottal vibration characteristics. Different locations of the constriction point along the vocal tract cause changes in the glottal vibration characteristics differently. Two variants of fricatives are examined, namely, alveolar fricative ([z]) and velar fricative ([ɤ]), which involve two different locations for the points of constriction of the vocal tract.

In the production of the apical lateral approximant ([l]) sound, the lateral stricture is relatively wide open for the entire steady-state duration [Fig. 1(d)], unlike that for [r] sound [Fig. 1(b)]. If the glottal vibration characteristics of the trill sounds are changed to normal modal vibration, then trills may sound like approximants (Mittal *et al.*, 2012). Apical lateral approximant ([l]) sounds are examined to understand the differences in the excitation characteristics from those of trills ([r]).

Nasal sounds involve closure at some location in the oral tract, while the nasal tract is kept open. Two variants of nasal sounds are examined, namely, alveolar nasal ([n]) and velar nasal ([ŋ]), to study whether the high stricture (nearly closed constriction) along the vocal tract, concurrent with the open nasal tract, has any effect on the glottal vibration.

Production of consonants ([r], [l], [z], [ɤ], [n], and [ŋ]) sounds in the context of vowel [a] are considered in this study. The sounds are only representative of a few sound categories. The *single*, *geminated*, and *prolonged* occurrences of these sounds are included in each category. The analysis of the effects of constriction is carried out using the *geminated* occurrence type for each of the six categories of sounds, as in this case the consonants can be produced in a sustained manner. The *single* cases are considered as these are the cases that usually occur in normal speech, and *prolonged* cases are studied to examine the effects due to prolongation.

III. SPEECH DATA FOR ANALYSIS

In natural production, speech sounds are produced as part of one or more syllables of the structure /CV/, /VCV/, or /VCCV/, consisting of vowels (/V/) and consonants (/C/). If the vowel on both sides is in modal voicing, then it is easier to distinguish the vowel and consonant regions for analysis. Consonants in the context of the open vowel [a] are considered in this study. Sometimes, changes in the production characteristics may not be highlighted in a single occurrence of consonant in the vowel context (/VCV/). Hence, sustained production of the consonants is considered. Sustained

production of consonants are either geminated (double) or prolonged (longer than geminated), i.e., in the form of /VCCV/ or /VCC…CV/ sound units, respectively. The distinctive characteristics of consonants may fade sometimes when they are prolonged. Hence, *geminated* type is used for detailed analysis.

Data were collected for the following six categories of voiced speech sounds: (1) Apical trill ([r]), (2) alveolar fricative ([z]), (3) velar fricative ([ɤ]), (4) apical lateral approximant ([l]), (5) alveolar nasal ([n]), and (6) velar nasal ([ŋ]). All these sounds are considered in the context of vowel [a] on both sides, in modal voicing. For each category of sound, three types of occurrences are considered: Single, geminated, and prolonged occurrence. Utterances of each type for each of the six categories were repeated three times. Thus the data consist of a total of 54 ($=6\times3\times3$) utterances. The data were collected in the voice of a male *expert* phonetician so as to have reliable and authentic data of production of these sounds. The data were also collected in the voice of a (less trained) female phonetics research student. Thus, total data have 108 ($=54+54$) utterances.

The data were recorded in a sound treated recording room. Simultaneous recordings of the speech signal and the EGG signal (Fourcin and Abberton, 1971; Fant, 1979; Fant *et al.*, 1985) were obtained for each utterance. The speech signal was recorded on a digital sound recorder with a high quality condenser microphone (Zoom H4n, Zoom Corp., Japan), kept at a distance of around 10 cm from the mouth. The EGG signal was recorded using an EGG recording device (Miller, 2012). The audio data were acquired at a sampling rate of 4400 samples/s, with 16 bits/sample. The data were downsampled to 10 000 samples/s before analysis. The collected data are available for download at the website of 'Speech and Vision Laboratory', IIIT, Hyderabad.

IV. FEATURES FROM SPEECH SIGNAL

A. Extraction of glottal excitation source characteristics

The features of the glottal source of excitation are derived from the speech signal using the ZFF method (Murty and Yegnanarayana, 2008; Yegnanarayana and Murty, 2009). In ZFF, the features of the impulse-like excitation of the glottal source are extracted by filtering the differenced speech signal through a cascade of two zero-frequency resonators (ZFRs). The key steps involved (Murty and Yegnanarayana, 2008) are as follows:

(a) A differenced speech signal $s[n]$ is considered. This preprocessing step removes the effect of any slow (low frequency) variations during recording of the signal, and produces a zero mean signal.

(b) The differenced signal $s[n]$ is passed through a cascade of two ZFRs, each of which is an all-pole system with two poles located at $z=+1$ in the z-plane. The output of the cascaded ZFRs is given by

$$y_1[n] = -\sum_{k=1}^{4} a_k y_1[n-k] + s[n], \tag{1}$$

where $a_1=-4$, $a_2=6$, $a_3=-4$, and $a_4=1$. It is equivalent to four successive cumulative sum (integration) operations in time-domain, which leads to a polynomial-like growth/decay of the ZFR output signal.

(c) The fluctuations in the ZFR output signal can be highlighted using a trend removal operation, which involves subtracting the local mean from the ZFR output signal at each time instant. The local mean is computed over a moving window of size $2N+1$ samples. The window size is about 1.5 times the average pitch period (in samples), which is computed using autocorrelation function of a 50 ms segment of the signal. The output of the trend removal operation is given by

$$y_2[n] = y_1[n] - \frac{1}{2N+1}\sum_{m=-N}^{N} y_1[n+m], \tag{2}$$

where $2N+1$ is the size of the window in number of samples. The resultant local mean subtracted signal is called the *zero-frequency filtered* signal. An illustration of the ZFF signal ($z_s[n]$) for a vowel segment is shown in Fig. 2(b), which is derived from the corresponding speech signal ($s[n]$) shown in Fig. 2(a).

(d) The positive to negative going zero-crossings correspond to the instants of glottal closure (GCIs), which are also referred to as *epochs* (Murty and Yegnanarayana, 2008). The interval between successive epochs corresponds to the fundamental period (T_0), inverse of which gives the instantaneous fundamental frequency (F_0) (Yegnanarayana and Murty, 2009).

(e) The slope of the ZFF signal around the epochs gives a measure of the *strength of* the impulse-like *excitation* (*SoE*) (Murty and Yegnanarayana, 2008; Yegnanarayana and Murty, 2009).

The *SoE* (denoted by ψ) at an epoch thus represents the relative amplitude of impulse-like excitation around that instant of significant excitation (i.e., GCI).

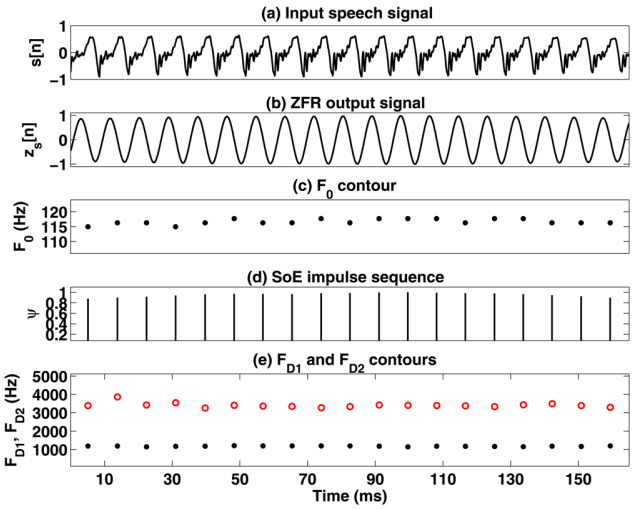


FIG. 2. (Color online) Illustration of waveforms of (a) input speech and (b) ZFF output signals, and (c) F_0 contour, (d) *SoE* values at epochs, (e) F_{D_1} ("●") and F_{D_2} ("○") derived from the speech signal for a vowel segment.

An illustration of F_0 contour and *SoE* impulse sequence is given in Figs. 2(c) and 2(d), respectively. The *SoE* impulse sequence represents the glottal source excitation, in which the location of each impulse corresponds to an epoch and its amplitude indicates relative strength of excitation around the epoch. The F_0 contour reflects the changes in successive epoch intervals.

B. Extraction of the vocal tract system characteristics

The vocal tract system characteristics are studied using the first two dominant frequencies F_{D_1} and F_{D_2} of the short-time spectral envelope. The features F_{D_1} and F_{D_2} of the vocal tract system are derived from the speech signal using LP spectrum (Makhoul, 1975). The steps involved are as follows:

(a) The vocal tract system characteristics are derived using LP analysis (Makhoul, 1975). Let $a_1, a_2, \ldots, a_p$ be the p LP coefficients. The corresponding all-pole filter $H(z)$ is given by

$$H(z) = \frac{1}{\left(1 - \sum_{k=1}^{p} a_k z^{-k}\right)}. \quad (3)$$

For a fifth order filter, there will be maximum of two peaks in the LP spectrum corresponding to two complex conjugate pole pairs. The frequencies corresponding to these peaks are called dominant peak frequencies, and are denoted as F_{D_1} and F_{D_2}.

(b) The group delay function $(\tau_g(\omega))$ is the negative derivative of the phase response of the all-pole filter (Murthy and Yegnanarayana, 1991), and is given by

$$\tau_g(\omega) = -\frac{d\theta(\omega)}{d\omega}, \quad (4)$$

where $\theta(\omega)$ is the phase angle of the frequency response $H(e^{jw})$ of the all-pole filter. The frequency locations of the peaks in the plot of the group delay function give the *dominant peak frequencies* (F_{D_1} and F_{D_2}).

The dominant peak frequencies F_{D_1} and F_{D_2} are derived for each pitch period using pitch synchronous LP analysis, anchored around GCIs. An illustration of F_{D_1} and F_{D_2} contours for the vowel segment in Fig. 2(a) is shown in Fig. 2(e).

C. Features for analysis

In this study, observations from the EGG and speech signals and the derived features are used for analysis of the six categories of sounds. Both qualitative and quantitative observations are discussed. Qualitative observations are made using waveforms of signals. Four waveforms are used for visual observation in each case: Speech signal, EGG signal, derivative of electroglottograph (dEGG), and ZFF output. Quantitative changes are measured from features extracted using speech signals, i.e., F_0, *SoE* (ψ), F_{D_1}, and F_{D_2}.

It is generally observed that changes in EGG, dEGG, and F_0 reflect the effect of glottal vibration, changes in F_{D_1} and F_{D_2} reflect the changes in the vocal tract system, and changes in the speech signal waveform, ZFF, and *SoE* may reflect changes in both the source and vocal tract system characteristics.

V. STUDY OF THE EFFECTS OF VOCAL TRACT CONSTRICTION USING EGG AND SPEECH SIGNALS

In this section the effects of vocal tract constriction in the production of different categories of sounds are examined in terms of observed and derived characteristics from EGG and speech signals. The effects caused by the size, type, and location of the stricture in the vocal tract are discussed in detail. The cross-sectional area of the opening at the stricture determines the *size* of the stricture, which in turn determines the *extent* of the (*high*, *low*, *no*) stricture. Very narrow to closed constriction in the vocal tract corresponds to *high* stricture, which occurs, for example, in trills ([r]) and alveolar fricatives ([z]), as in Figs. 1(b) and 1(c), respectively. The intermediate case of a relatively wider opening corresponds to *low* stricture, which occurs in the case of the approximant ([l]) [Fig. 1(d)] and the velar fricative ([ɤ]). A completely open vocal tract corresponds to *no* stricture, as in the case of open vowel [a].

Two different *types* of strictures are considered in this study: *cyclic* {as in [r] in Fig. 1(b)} and *steady* {as for [z] in Fig. 1(c) and [l] in Fig. 1(d)}. Two different *locations* of strictures in the vocal tract are considered, namely *alveolar* (as in [z]) and *velar* (as in [ɤ]). In addition, the effects of closed vocal tract (*high* stricture) during production of nasal sounds are considered for any possible effect in terms of the observed and derived characteristics from EGG and speech signals. Two different locations of the stricture for nasals are considered, namely, alveolar nasal ([n]) and velar nasal ([ŋ]). All these categories of sounds are produced in the context of the open vowel [a], where there is *no* stricture. Only geminated utterances of these different categories of sounds are analyzed in this section, as gemination of consonants produces sustained output that facilitates study of the effects of constriction, while eliminating possible effects of vowel-consonant transition.

Figures 3–8 show the waveforms of speech and EGG signals for the six categories of sounds chosen for analysis in this section. Each figure displays, besides the waveforms of speech, EGG, dEGG, and ZFF output signals, the contours of instantaneous fundamental frequency (F_0), strength of excitation (*SoE*), and the dominant peak frequencies (F_{D_1} and F_{D_2}).

A. Apical trill ([r])

In the production of apical trill ([r]) sound, the *high* stricture is formed due to a narrow opening between the alveolar/palatal region and the apical region of the tongue. This stricture gets broken, releasing the air pressure built in the oral cavity, and it is formed again due to the Bernoulli effect (Catford, 2001). Thus this stricture is *cyclic* in nature, due to opening and closing of the stricture in each trill cycle [Fig. 1(b)].

The cyclic high stricture affects the rate of vibration of the vocal folds and the strength of excitation (Dhananjaya

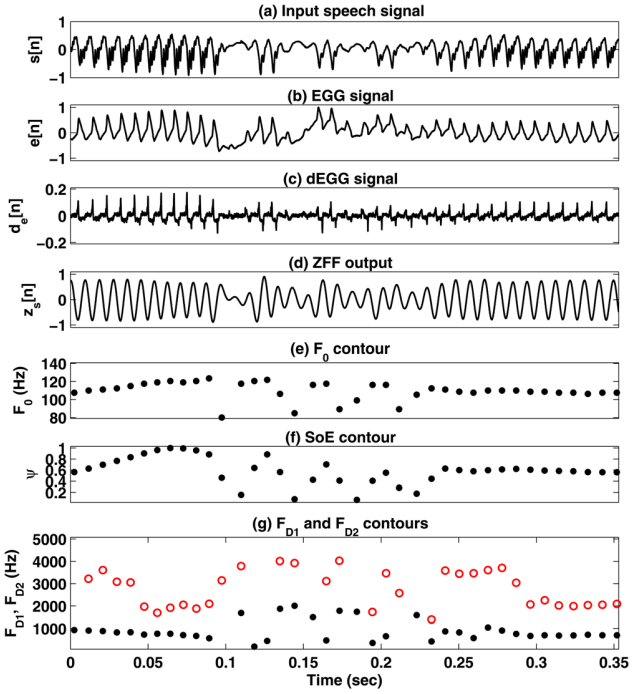


FIG. 3. (Color online) Illustration of waveforms of (a) input speech signal, (b) EGG signal, (c) dEGG signal, (d) ZFF output, and features (e) F_0, (f) *SoE*, (g) F_{D_1} ("●") and F_{D_2} ("○") for geminated occurrence of *apical trill* ([r]) in the vowel context [a]. The sound is for [arra], produced in male voice.

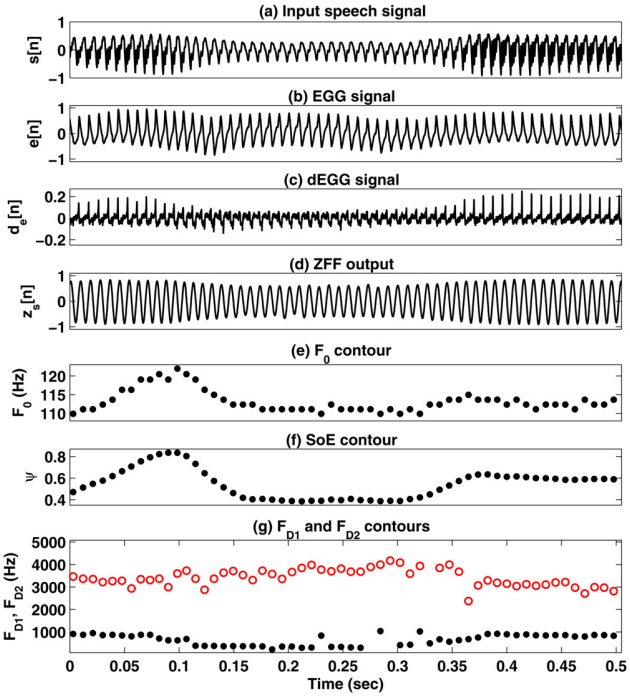


FIG. 5. (Color online) Illustration of waveforms of (a) input speech signal, (b) EGG signal, (c) dEGG signal, (d) ZFF output, and features (e) F_0, (f) *SoE*, (g) F_{D_1} ("●") and F_{D_1} ("○") for geminated occurrence of *velar fricative* ([ɤ]) in the vowel context [a]. The sound is for [aɤɤa], produced in male voice.

et al., 2012). These effects are reflected in the F_0 and *SoE* contours in Figs. 3(e) and 3(f), respectively. This is because the pressure gradient across the glottis reduces during the closed phase of the trill cycle, which in turn reduces F_0.

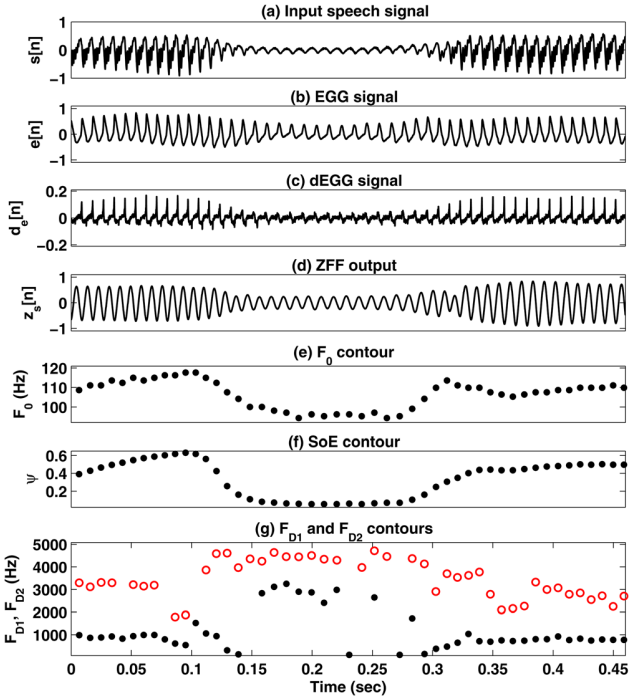


FIG. 4. (Color online) Illustration of waveforms of (a) input speech signal, (b) EGG signal, (c) dEGG signal, (d) ZFF output, and features (e) F_0, (f) *SoE*, (g) F_{D_1} ("●") and F_{D_2} ("○") for geminated occurrence of *alveolar fricative* ([z]) in the vowel context [a]. The sound is for [azza], produced in male voice.

Thus the coupling effect of the system on the source is significant in this case.

There are also changes in the resonances of the vocal tract system due to changes in the shape of the tract during

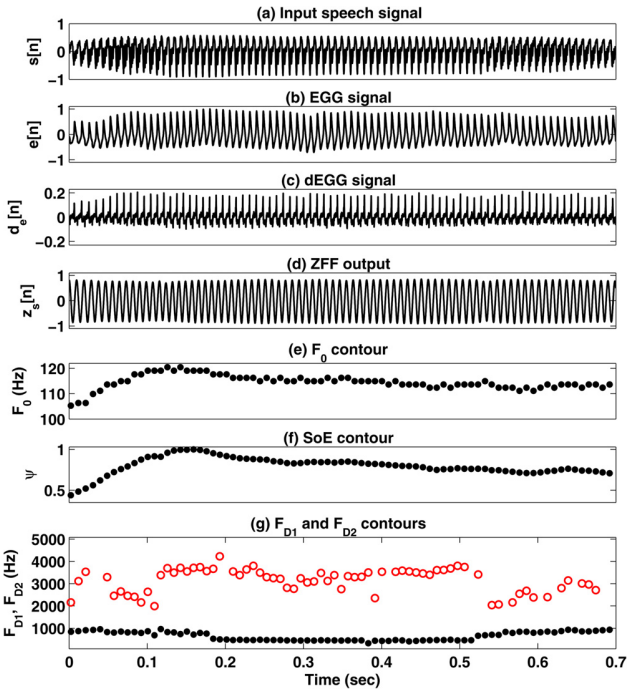


FIG. 6. (Color online) Illustration of waveforms of (a) input speech signal, (b) EGG signal, (c) dEGG signal, (d) ZFF output, and features (e) F_0, (f) *SoE*, (g) F_{D_1} ("●") and F_{D_2} ("○") for geminated occurrence of *apical lateral approximant* ([l]) in the vowel context [a]. The sound is for [alla], produced in male voice.

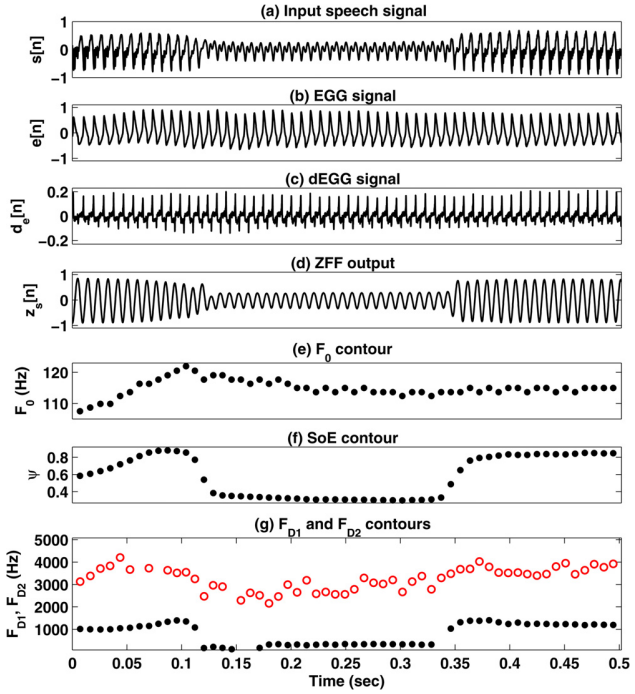


FIG. 7. (Color online) Illustration of waveforms of (a) input speech signal, (b) EGG signal, (c) dEGG signal, (d) ZFF output, and features (e) F_0, (f) *SoE*, (g) F_{D_1} ("●") and F_{D_2} ("○") for geminated occurrence of *alveolar nasal* ([n]) in the vowel context [a]. The sound is for [anna], produced in male voice.

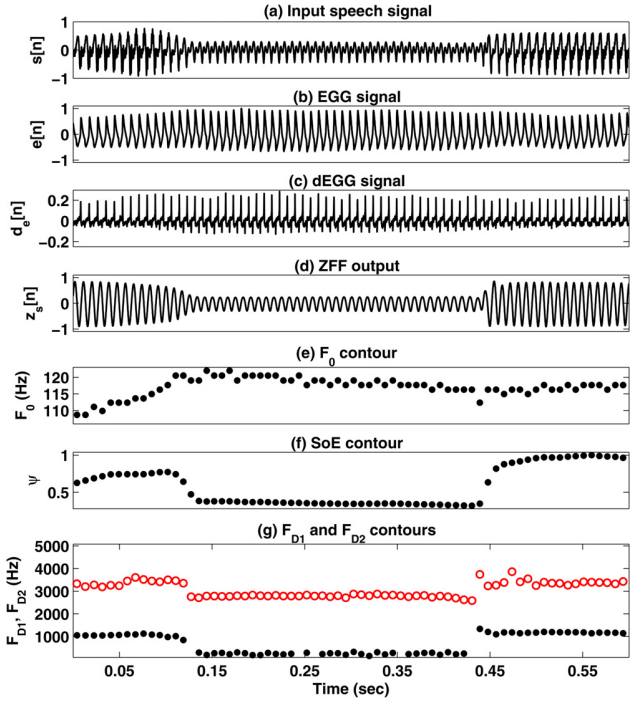


FIG. 8. (Color online) Illustration of waveforms of (a) input speech signal, (b) EGG signal, (c) dEGG signal, (d) ZFF output, and features (e) F_0, (f) *SoE*, (g) F_{D_1} ("●") and F_{D_2} ("○") for geminated occurrence of *velar nasal* ([ŋ]) in the vowel context [a]. The sound is for [aŋŋa], produced in male voice.

each trill cycle. These changes are seen as cyclic variations of F_{D_1} and F_{D_2} [Fig. 3(g)], where F_{D_1} is higher during the closed phase of the trill cycle.

The effects of constriction due to dynamic vocal tract configuration can be seen in the waveforms of EGG, dEGG, and the speech signal (Fig. 3). The contrast between the steady vowel region and the trill region can be seen in all signals and in the features derived from the signals.

B. Alveolar fricative ([z])

Production of alveolar fricative ([z]) involves a narrow opening of the constriction between the upper articulator (alveolar ridge) and the lower articulator (tongue tip), as shown in Fig. 1(c). The constriction is narrow enough to produce frication or turbulence. Thus [z] is produced by a *high steady* stricture, unlike the *high cyclic* stricture in the case of [r]. There is pressure buildup behind the constriction, causing pressure differential across the glottis. Thus in this case, the constriction effect can be seen in the signal waveform, as well as in the source and system features derived from the signals.

Amplitudes of the speech signal, EGG, dEGG, and ZFF are low, relative to the adjacent vowel (Fig. 4). The constriction results in lowering of F_0 and *SoE* values, relative to the adjacent vowel region, as can be seen in Figs. 4(e) and 4(f), respectively. Due to frication, both the dominant peak frequencies (F_{D_1} and F_{D_2}) show high values, compared to those in the vowel region [Fig. 4(g)]. The constriction effects are similar to the trill case ([r]), except that in the case of alveolar fricative ([z]) the effects are steady, and not cyclic.

C. Velar fricative ([ɤ])

The production of velar fricative ([ɤ]) involves *steady* but relatively *lower* stricture due to more opening in the constriction between the upper and lower articulators, than for the alveolar fricative ([z]). Since the constriction area has to be small enough to produce turbulence, this stricture may be termed as *steady high-low* stricture, and the constriction effects are expected to be similar to those for the alveolar fricatives.

While there are no significant changes in the EGG signal waveform, relative to the adjacent vowel region [Fig. 5(b)], changes can be seen better in the waveform of dEGG signal [Fig. 5(c)]. Constriction effects can be seen in the derived source information, i.e., F_0 contour [Fig. 5(e)] and *SoE* contour [Fig. 5(f)], whereas the changes are less evident in the ZFF signal [Fig. 5(d)].

The changes in the speech signal waveform for velar fricative ([ɤ]) relative to that for vowel [a] can be attributed to the changes in the vocal tract characteristics. Turbulence generated at the stricture is lower in the case of [ɤ], than in the case of [z]. As a result, the F_{D_1} is lower than for the vowel [a] [Fig. 5(g)]. Also, the frication effect is not as high as in the case of [z], and hence the behavior of F_{D_1} and F_{D_2} are more vowel-like, in the sense that they are in the same range as that for the vowel [a] [Fig. 5(g)], unlike for [z] where both F_{D_1} and F_{D_2} are high [Fig. 4(g)].

D. Approximant ([l])

An apical lateral approximant ([l]) is formed by a closure between the alveolar/palatal region (upper articulator) and the apical tongue region (lower articulator), along with a simultaneous lateral stricture. In this case, the lateral stricture is wide enough, as shown in Fig. 1(d), to allow free flow of air in the vocal tract. Thus the stricture is *low*, i.e., relatively more open than for the *high* stricture cases considered so far ([r] and [z]), and is also *steady*, i.e., not cyclic as in the case of trill [r]. Since there is no significant pressure gradient buildup in this case, the constriction effect is negligible in comparison with the *high* stricture case. One does not notice any significant changes in the amplitudes in the waveforms of speech signal, EGG, dEGG, and ZFF output, relative to the adjacent vowel regions, as can be seen in Figs. 6(a)–6(d).

There are no major changes in the excitation features as well [see F_0 and *SoE* contours in Figs. 6(e) and 6(f), respectively]. However, due to a wider lateral opening, the corresponding change in the shape of the vocal tract affects the first two dominant peak frequencies [Fig. 6(g)]. The F_{D_1} is reduced and F_{D_2} is increased, relative to the values in the neighboring vowel region. This shows that if the stricture is not high, the constriction effects are negligible.

E. Alveolar nasal ([n]) and velar nasal ([ŋ])

Nasal sounds are produced with complete closure of the vocal tract at some location in the oral cavity, and simultaneous flow of air through the nasal tract, which is facilitated by the velic opening. Here, the constriction along the vocal tract is like a *high* stricture case, but due to the coupling of the nasal tract there is hardly any obstruction to the egressive flow of air. Nasal sounds are considered to examine whether the *high* stricture in the vocal tract has any effect on the glottal excitation. Two variants of the high stricture along the vocal tract are considered, corresponding to two different locations, namely, alveolar nasal ([n]) and velar nasal ([ŋ]).

Figures 7 and 8 show the waveforms and other features for [n] and [ŋ], respectively. Due to the absence of constriction effect on the glottal vibration, there are no visible changes in EGG and dEGG waveforms, in relation to the adjacent vowel. Also, there is hardly any significant change in the F_0 contours [Figs. 7(e) and 8(e)], indicating that the glottal vibration is not affected.

However, there is reduction in the amplitude of the waveform of the speech signal in both cases of [n] and [ŋ], as shown in Figs. 7(a) and 8(a), respectively. This is primarily due to a narrow constricted (turbinated) path of the nasal tract, especially at the nares. The effect of this constriction can also be seen in the significantly lower amplitudes of ZFF output [Figs. 7(d) and 7(f)] and *SoE* contour [Figs. 8(d) and 8(f)], as compared to the adjacent vowel [a]. As expected, the resonance frequency due to nasal tract coupling is significantly lower than for the vowel, as can be seen in the F_{D_1} contours in Figs. 7(g) and 8(g), for [n] and [ŋ], respectively. In fact, even the F_{D_2} is also lower in both cases, although this change is more clearly visible in the case of [ŋ] [Fig. 8(g)].

In summary, in the case of nasal sounds ([n] and [ŋ]), the high stricture in the vocal tract does not cause any constriction effect on the glottal excitation. However, there are significant changes in the speech signal, ZFF signal, *SoE*, F_{D_1}, and F_{D_2}, relative to the adjacent vowel. These changes are primarily due to narrow constriction in the nasal tract.

In Table I, comparisons among different sound categories are made, based on the level of stricture in the vocal tract. In each case, the difference in the stricture size, type, and location in the vocal tract that causes the difference in the constriction effect, is highlighted. The signal waveforms and the derived features that are mostly/sometimes/not affected for each sound type are marked to provide a complete picture of the effects of constriction of the vocal tract.

VI. QUANTITATIVE ASSESSMENT OF THE EFFECTS OF VOCAL TRACT CONSTRICTION

The effects of constriction were examined for geminated cases of the six sound categories, where the production of sound is sustained. It would be interesting to observe changes in the features for single and prolonged occurrences

TABLE I. Comparison between sound types based on stricture differences for geminated occurrences. *Abbreviations:* alfric: alveolar fricative [z], vefric: velar fricative [ɤ], approx/appx: approximant [l], frics: fricatives ([z], [ɤ]), alnasal: alveolar nasal [n], venas: velar nasal [ŋ], stric: stricture, H/L indicates relative degree of low stricture.

#	Categories of sounds	Main causes for difference in effects of constriction	Qualitative observations (using waveforms)				Quantitative observations (using features)			
			(a) $s[n]$	(b) $e[n]$	(c) $d_e[n]$	(d) $z_s[n]$	(e) F_0	(f) ψ	(g) F_{D1}	(h) F_{D2}
1.	trill *vs* vowel ([r] *vs* [a])	cyclic high stric:[r] steady no stric:[a]	●	●	●	●	✓	✓	✓	●
2.	alfric *vs* trill ([z] *vs* [r])	steady high stric:[z] cyclic high stric:[r]	✓	●	✓	✓	✓	✓	✓	●
3.	vefric *vs* alfric ([ɤ] *vs* [z])	steady Hlow stric:[ɤ] steady high stric:[z]	●	●	✓	●	●	○	✓	○
4.	approx *vs* vefric ([l] *vs* [ɤ])	steady low stric:[l] steady Hlow stric:[ɤ]	●	○	●	○	●	✓	●	✓
5.	approx *vs* vowel ([l] *vs* [a])	steady low stric:[l] steady no stric:[a]	●	○	○	○	○	○	✓	●
6.	trill *vs* approx ([r] *vs* [l])	cyclic high stric:[r] steady low stric:[l]	●	✓	✓	●	✓	✓	✓	●
7.	frics *vs* trill/appx ([z],[ɤ] *vs* [r],[l])	high stric:[z],[r] H/L low stric:[ɤ],[l]	●	●	✓	●	●	✓	✓	○
8.	nasals *vs* vowel ([n], [ŋ] *vs* [a])	nasal low stnc:[n], [ŋ] steady no stric:[a]	✓	○	●	✓	○	✓	✓	●
9.	nasals *vs* approx ([n], [ŋ] *vs* [l])	nasal low stric:[n], [ŋ] steady low stric:[l]	●	○	○	●	○	✓	●	✓
10	alnasal *vs* venas ([n] *vs* [ŋ])	nasal high stric:[n], nasal Hlow stric:[ŋ]	○	○	●	○	○	○	✓	●

Legend:- ✓: mostly evident, ●: sometimes/less evident, ○: rarely/not evident changes

of these sounds, relative to the geminated occurrences. In this section, changes in the vibration characteristics due to vocal tract constriction are examined using the average values of the features.

Average values of the features F_0, *SoE*, F_{D_1}, and F_{D_2} are obtained in the regions of consonant and the vowel for each case, and are given in Tables II and III. The average values of each sound category over the three types of occurrences (single, geminated, and prolonged) are given in Tables IV and V.

Changes in F_0 and *SoE* in comparison to those for the vowel [a] are given in Table II. The average values of F_0 for vowel [a], minimum F_0 and maximum F_0 (i.e., $F_{0_{[a]}}$, $F_{0_{\min}}$, and $F_{0_{\max}}$) for each sound category are given (in Hz) in columns (a), (b), and (c), respectively. The values are rounded off to a single decimal. The percentage change in F_0 relative to $F_{0_{[a]}}$, i.e., $\Delta F_0/F_{0_{[a]}} (= (F_{0_{\max}} - F_{0_{\min}}/F_{0_{[a]}})\%)$ is given in column (d). Likewise, the values of the feature *SoE* (denoted as ψ) are given in columns (e)–(h). The features are normalized relative to those for the vowel (i.e., $F_{0_{[a]}}$ and $\psi_{[a]}$), to facilitate comparison across sound categories. Since the number of points (GCIs) for feature values is less in some cases of sounds such as single or geminated occurrences of trill ([r]) sounds, the range of deviation in the feature is computed using minimum and maximum values, rather than computing the standard deviation.

Significant changes in F_0 and *SoE* in comparison to the vowel [a] can be observed for apical trill ([t]) and alveolar fricative ([z]), in columns (d) and (h) in Table II. A dip in F_0 for alveolar fricative ([z]) is due to constriction of the nearly closed vocal tract (to produce fricative noise for sound [z]) on the vibration of the vocal folds. The strength of excitation (ψ) is also reduced for [z] due to constriction in the vocal tract. A sharp change (a dip) in *SoE* for both nasals ([n] and [ŋ]) can be observed from column (h). This is because of the constriction in the nasal tract, and not due to glottal vibration, as in the case for [r] or [z]. Absence of the effect of vocal tract constriction on excitation is evident from the negligible changes in the F_0 values in the case of nasals.

In Table III, the average values of F_{D_1} for the vowel [a] ($F_{D_{1_{[a]}}}$), $F_{D_{1_{\min}}}$ (minimum F_{D_1}) and $F_{D_{1_{\max}}}$ (maximum F_{D_1}) are given in columns (a), (b), and (c), respectively. Percentage changes in F_{D_1} for these sounds relative to $F_{D_{1_{[a]}}}$ (for vowel), i.e., $\Delta F_{D_1}/F_{D_{1_{/a/}}} (= (F_{D_{1_{\max}}} - F_{D_{1_{\min}}}/F_{0_{[a]}})\%)$, are given in column (d). Likewise, the average values of F_{D_2} are given in columns (e)–(h). The values of F_{D_1} and F_{D_2} are rounded off to the nearest integers, and the percentage changes to a single decimal. Large changes can be observed in F_{D_1} for trill [r] and alveolar fricative [z], relative to the vowel [a]. In comparison, the changes in F_{D_1} for velar fricative ([ɤ]) and lateral approximant ([l]) are relatively low. The changes in F_{D_1} for nasals ([n] and [ŋ]) are significant, due to lowering of the first formant of the nasal tract. Changes in F_{D_2} are high mainly for the trill ([r]) sound, as can be seen in column (h).

A summary of percentage changes in F_0, *SoE*, F_{D_1}, and F_{D_2}, relative to those for vowel [a], for the six categories of sounds for the *male voice*, is given in Table IV. The average values of the changes in these feature computed across the three types of occurrences are given for each sound category. In each case, the relative increase or decrease (i.e., the direction of change) in the average values of these features, in comparison to those for the vowel [a], is marked as (+) or (−), respectively. Significant changes in F_0 are due to the effect of vocal tract constriction on the glottal vibration, and in F_{D_1} due to changes in the system characteristics. The summary table clearly illustrates the changes in different sound categories as discussed before.

The summary of changes in features F_0, *SoE*, F_{D_1}, and F_{D_2} for a subset of the data collected for a female voice is given in Table V in columns (a)–(d), respectively.

TABLE II. Changes in glottal source features F_0 and *SoE* (ψ) for six categories of sounds (for *male* voice). Column (a) is F_0 (Hz) for vowel [a], (b) and (c) are $F_{0_{\min}}$ and $F_{0_{\max}}$ for the specific sound, and (d) is $\Delta F_0/F_{0_{[a]}}(\%)$. Column (e) is *SoE* (i.e., ψ) for vowel [a], columns (f) and (g) are $\psi_{\min}$ and $\psi_{\max}$ for the specific sound, and column (h) is $\Delta\psi/\psi_{[a]}(\%)$. Suffixes a, b, and c in the first column indicate single, geminated, or prolonged occurrences, respectively. Note: "alfric"/"vefric" denotes alveolar/velar fricative and "alnasal"/"venasal" denotes alveolar/velar nasal.

Sl. #	Sound category	Sound Symbol	(a) $F_{0_{[a]}}$	(b) $F_{0_{\min}}$	(c) $F_{0_{\max}}$	(d)(%) $\Delta F_0/F_{0_{[a]}}$	(e) $\psi_{[a]}$	(f) $\psi_{\min}$	(g) $\psi_{\max}$	(h)(%) $\Delta\psi/\psi_{[a]}$
1a	trill	[ara]	112.1	88.5	117.7	26.00	0.820	0.237	0.987	91.34
1b	trill	[arra]	111.1	85.8	118.3	29.26	0.665	0.119	0.753	95.31
1c	trill	[arr…ra]	111.4	89.4	118.5	26.06	0.734	0.146	0.634	66.50
2a	alfric	[aza]	111.4	95.9	117.0	18.91	0.617	0.074	0.751	109.6
2b	alfric	[azza]	110.6	94.6	116.3	19.59	0.509	0.057	0.566	99.96
2c	alfric	[azz…za]	111.5	95.6	117.7	19.85	0.641	0.075	0.740	103.8
3a	vefric	[aɤa]	112.7	111.1	117.7	5.81	0.813	0.627	0.943	38.90
3b	vefric	[aɤɤa]	112.2	110.9	119.1	7.34	0.608	0.393	0.735	56.23
3c	vefric	[aɤɤ…ɤa]	112.1	111.7	117.7	5.30	0.798	0.538	0.957	52.57
4a	lateral	[ala]	114.9	117.6	119.1	1.22	0.787	0.889	0.933	5.67
4b	lateral	[alla]	113.3	114.0	119.8	5.11	0.720	0.819	0.903	11.64
4c	lateral	[all…la]	114.6	112.8	120.1	6.42	0.616	0.582	0.707	20.24
5a	alnasal	[ana]	112.7	114.7	118.6	3.48	0.744	0.299	0.814	69.09
5b	alnasal	[anna]	113.0	113.4	119.1	5.08	0.748	0.304	0.861	74.39
5c	alnasal	[ann…na]	115.7	113.2	117.7	3.85	0.793	0.251	0.813	70.81
6a	venasal	[aŋa]	112.7	114.3	119.0	4.15	0.722	0.311	0.862	76.30
6b	venasal	[aŋŋa]	114.8	115.0	119.8	4.20	0.828	0.331	0.879	66.14
6c	venasal	[aŋŋ…ŋa]	115.6	117.1	119.9	2.46	0.748	0.315	0.880	75.59

TABLE III. Changes in vocal tract system features F_{D_1} and F_{D_2} for six categories of sounds (for *male* voice). Column (a) is F_{D_1} (Hz) for vowel [a], (b) and (c) are $F_{D1_{\min}}$ and $F_{D1_{\max}}$ for the specific sound, and (d) is $\Delta F_{D1}/F_{D1_{[a]}}(\%)$. Column (e) is F_{D_2} (Hz) for vowel [a], columns (f) and (g) are $F_{D2_{\min}}$ and $F_{D2_{\max}}$ for the specific sound, and column (h) is $\Delta F_{D2}/F_{D2_{[a]}}(\%)$. Suffixes a, b, and c in the first column indicate single, geminated, or prolonged occurrences, respectively. Note: "alfric"/"vefric" denotes alveolar/velar fricative and "alnasal"/"venasal" denotes alveolar/velar nasal.

Sl. #	Sound category	Sound Symbol	(a) $F_{D_{1_{[a]}}}$	(b) $F_{D_{1_{\min}}}$	(c) $F_{D_{1_{\max}}}$	(d)(%) $\Delta F_{D_1}/F_{D_{1_{[a]}}}$	(e) $F_{D_{2_{[a]}}}$	(f) $F_{D_{2_{\min}}}$	(g) $F_{D_{2_{\max}}}$	(h)(%) $\Delta F_{D_2}/F_{D_{2_{[a]}}}$
1a	trill	[ara]	761	525	1499	128.1	2022	1377	3506	105.3
1b	trill	[arra]	763	402	1837	188.0	2006	1655	3933	113.6
1c	trill	[arr…ra]	791	397	1882	187.7	2399	1863	3793	80.5
2a	alfric	[aza]	856	278	2723	285.6	2892	3612	4451	29.0
2b	alfric	[azza]	844	227	2873	313.6	3092	3678	4538	27.8
2c	alfric	[azz…za]	886	288	2749	277.7	3250	3875	4536	20.4
3a	vefric	[aɤa]	735	363	913	74.9	3195	3245	3752	15.9
3b	vefric	[aɤɤa]	867	342	968	72.2	3131	3062	3884	26.3
3c	vefric	[aɤɤ…ɤa]	889	359	1031	75.6	3114	3182	3855	21.6
4a	lateral	[ala]	804	562	732	21.3	2301	1748	3725	85.9
4b	lateral	[alla]	862	480	652	20.0	2361	2349	4229	79.6
4c	lateral	[all…la]	815	361	495	16.4	2774	2589	3746	41.7
5a	alnasal	[ana]	1137	327	1410	95.3	3483	2454	4010	44.7
5b	alnasal	[anna]	1103	241	1387	103.9	3440	2491	3523	30.0
5c	alnasal	[ann…na]	1084	267	1240	89.7	3513	2842	3694	24.3
6a	venasal	[aŋa]	1177	261	1316	89.6	3335	2438	3925	44.6
6b	venasal	[aŋŋa]	1119	244	1149	80.8	3277	2778	3650	26.6
6c	venasal	[aŋŋ…ŋa]	1118	214	1195	87.7	3329	2713	3689	29.3

Tables IV and V show some differences. The extent of changes in F_0, *SoE*, F_{D_1}, and F_{D_2} for [r] seem to be less for the female speaker. The signs of changes in *SoE* ($\Delta\psi$) for both nasals ([n] and [ŋ]), as compared to the vowel [a], are also different for both the speakers. It could possibly be related to a higher average pitch of the female voice in comparison to the male voice. Another reason for the differences in Tables IV and V could be that the data for Table IV were obtained from an expert phonetician, whereas the data for Table V were obtained from a research scholar with only basic training in phonetics. It is likely that the female speaker could not articulate some of the sounds clearly.

VII. SUMMARY AND CONCLUSIONS

In this study, the effects of constriction of the vocal tract system on the vibration characteristics of the vocal folds are examined for different types of constrictions that occur in the production of some speech sounds. Involuntary changes in glottal vibrations in the production of some specific categories of sounds are examined. Six categories of sounds are considered for illustration, which differ in the size, type, and location of stricture in the vocal tract. The study considers sounds uttered in modal voicing, in the context of vowel [a]. Single, geminated, and prolonged occurrences are examined for each sound category. We have considered features describing the glottal vibration and the vocal tract system to demonstrate the effect of system-source coupling. Changes due to different types of constrictions are observed in the amplitudes of the waveforms of speech signal, EGG, dEGG, and ZFF output, and also in the features F_0, *SoE*, F_{D_1}, and F_{D_2}. The glottal source features F_0 and *SoE* are derived from the speech signal using the ZFF method. The vocal tract system characteristics are represented through two dominant peak frequencies F_{D_1} and F_{D_2}, which are derived by LP analysis of the speech signal.

The general observation is that a high degree of constriction causing obstruction to the flow of air results in large changes in F_0 and strength of impulse-like excitation,

TABLE IV. Changes in features due to the effects of vocal tract constriction on the glottal vibration, for six categories of sounds (*male* voice). Columns (a)–(d) show percentage changes in F_0, *SoE*, F_{D_1}, and F_{D_2}, respectively. The direction of change in a feature in comparison to that for vowel [a] is marked with ±sign. Note: "alfric"/"vefric" denotes alveolar/velar fricative and "alnasal"/"venasal" denotes alveolar/velar nasal.

Sl. #	Sound category	IPA Symbol	(a)(%) ΔF_0	(b)(%) $\Delta\psi$	(c)(%) ΔF_{D_1}	(d)(%) ΔF_{D_2}
1	trill	[r]	−27.1	−84.8	+167.9	+99.8
2	alfric	[z]	−19.5	−104.5	+292.3	+25.7
3	vefric	[ɤ]	+6.1	−49.2	−74.2	+21.2
4	lateral	[l]	+4.3	+12.5	−19.2	+69.1
5	alnasal	[n]	+4.1	−71.4	−96.3	−33.0
6	venasal	[ŋ]	+3.6	−72.7	−86.0	−33.5

TABLE V. Changes in features due to the effects of vocal tract constriction on the glottal vibration, for six categories of sounds (*female* voice). Columns (a)–(d) show percentage changes in F_0, *SoE*, F_{D_1}, and F_{D_2}, respectively. The direction of change in a feature in comparison to that for vowel [a] is marked with ±sign. Note: "alfric"/"vefric" denotes alveolar/velar fricative and "alnasal"/"venasal" denotes alveolar/velar nasal.

Sl. #	Sound category	IPA Symbol	(a)(%) ΔF_0	(b)(%) $\Delta\psi$	(c)(%) ΔF_{D_1}	(d)(%) ΔF_{D_2}
1	trill	[r]	+15.5	−49.1	+72.4	+30.9
2	alfric	[z]	−29.4	−116.2	+273.3	+43.8
3	lateral	[l]	+2.9	+35.0	−51.7	+17.1
4	alnasal	[n]	+6.6	+46.4	−108.7	−21.0
5	venasal	[ŋ]	+6.6	+22.5	−72.4	−26.3

relative to the values in the adjacent steady vowel regions. This happens in the cases of apical trill ([r]) and alveolar fricative ([z]) sounds. The stricture in the vocal tract is *high* (i.e., narrow constriction) in these cases. There is also a significant increase in the first dominant peak frequency (F_{D_1}) in the spectrum. Changes in these features are less significant in the cases when the obstruction to the airflow is *low*, as in the cases of velar fricative ([ɤ]) and lateral approximant ([l]) sounds. If there is *no* obstruction to the airflow as in the case of nasal sounds ([n] and [ŋ]), then there is hardly any change in the characteristics of the glottal vibration in relation to the adjacent vowel. Note that in the case of nasals the strength of excitation is reduced due to constriction in the nasal tract. Associated changes (reduction) in the dominant peak frequencies F_{D_1} and F_{D_2} are primarily due to increasing effective length caused by coupling of the nasal tract. Thus this study shows that the features of the glottal vibration such as F_0 and strength of excitation as well as the dominant peak frequencies, all of which can be derived from the signal, can be used to infer the degree of constriction in the vocal tract during production of speech sounds.

The study examines the nature of involuntary changes in the glottal vibration characteristics due to the vocal tract constriction, along with associated changes in the vocal tract system characteristics. Only a select set of sounds are examined in this study. More variety of sounds and their variants need to be studied further. Also, the effects of different vowel contexts need to be examined. Different features may also be needed to understand the differences in the characteristics of sounds from production point of view.

Barney, A., Shadle, C. H., and Davies, P. O. A. L. (**1999**). "Fluid flow in a dynamic mechanical model of the vocal folds and tract. I. Measurements and theory," J. Acoust. Soc. Am. **105**, 444–455.

Barney, A., Stefano, A. D., and Henrich, N. (**2007**). "The effect of glottal opening on the acoustic response of the vocal tract," Acta Acust. Acust. **93**, 1046–1056.

Catford, J. C. (**2001**). *A Practical Introduction to Phonetics*, 2nd ed. (Oxford University Press Inc., New York), Chap. 4, pp. 59–69.

Chan, R. W., and Titze, I. R. (**2006**). "Dependence of phonation threshold pressure on vocal tract acoustics and vocal fold tissue mechanics," J. Acoust. Soc. Am. **119**, 2351–2362.

Chi, X., and Sonderegger, M. (**2007**). "Subglottal coupling and its influence on vowel formants," J. Acoust. Soc. Am. **122**, 1735–1745.

Dhananjaya, N., Yegnanarayana, B., and Bhaskararao, P. (**2012**). "Acoustic analysis of trill sounds," J. Acoust. Soc. Am. **131**, 3141–3152.

Ewan, W. G. (**1977**). "Can the intrinsic F_0 differences between vowels be explained by source/tract coupling?," Status Report on Speech Research, Haskins Laboratories SR-51/52, pp. 197–199.

Ewan, W. G., and Ohala, J. J. (**1979**). "Can intrinsic vowel F_0 be explained by source/tract coupling?," J. Acoust. Soc. Am. **66**, 358–362.

Fant, G. (**1970**). *Acoustic Theory of Speech Production*, 1st ed. (Mounton Co. N. N. Publishers, The Hague, Netherlands), Chap. 1.1, pp. 15–24.

Fant, G. (**1979**). "Glottal source and excitation analysis," Speech Transmission Laboratory, KTH, Sweden, Quarterly Progress and Status Report 20, pp. 85–107.

Fant, G. (**2004**). *Speech Acoustics and Phonetics Selected Writings*, Text, Speech and Language Technology, Vol. 24, 1st ed. (Kluwer Academic Publishers, Dordrecht, The Netherlands), Chap. 4.1, pp. 143–161.

Fant, G., and Lin, Q. (**1987**). "Glottal source—vocal tract acoustic interaction," Speech Transmission Laboratory, KTH, Sweden, Quarterly Progress and Status Report 28, pp. 13–27.

Fant, G., Lin, Q., and Gobl, C. (**1985**). "Notes on glottal flow interaction," Speech Transmission Laboratory, KTH, Sweden, Quarterly Progress and Status Report 26, pp. 21–45.

Fourcin, A., and Abberton, E. (**1971**). "First application of a new laryngograph," Med. Biol. Illus. **21**, 172–182.

Hatzikirou, H., Fitch, W. T., and Herzel, H. (**2006**). "Voice instabilities due to source-tract interactions," Acta Acust. Acust. **92**, 468–475.

Ladefoged, P., and Johnson, K. (**2011**). *A Course in Phonetics*, 6th ed. (Cengage Learning India Private Limited, Delhi, India), Chap. 1, pp. 4–7.

Laver, J. (**1994**). *Principles of Phonetics*, Cambridge Textbooks in Linguistics (Cambridge University Press, London), Chap. 5, pp. 119–158.

Lucero, J. C., Lourenco, K. G., Hermant, N., Hirtum, A. V., and Pelorson, X. (**2012**). "Effect of source-tract acoustical coupling on the oscillation onset of the vocal folds," J. Acoust. Soc. Am. **132**, 403–411.

Makhoul, J. (**1975**). "Linear prediction: A tutorial review," Proc. IEEE **63**, 561–580.

McGowan, R. S. (**1992**). "Tongue-tip trills and vocal-tract wall compliance," J. Acoust. Soc. Am. **91**, 2903–2910.

Miller, D. G. (**2012**). "EGGs for singers," URL http://www.eggsforsingers.eu/ (Last viewed April 1, 2013).

Mittal, V. K., Dhananjaya, N., and Yegnanarayana, B. (**2012**). "Effect of tongue tip trilling on the glottal excitation source," in *Proceedings of INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association*, Portland, Oregon.

Murthy, H. A., and Yegnanarayana, B. (**1991**). "Formant extraction from group delay function," Speech Commun. **10**, 209–221.

Murty, K. S. R., and Yegnanarayana, B. (**2008**). "Epoch extraction from speech signals," IEEE Trans. Audio, Speech, Lang. Process. **16**, 1602–1613.

Ohala, J. J., and Eukel, B. W. (**1987**). "Explaining the intrinsic pitch of vowels," Channon, Shockey, in Honor of Ilse Lehiste, pp. 207–215.

Perkell, J. S., and Cohen, M. H. (**1989**). "An indirect test of the quantal nature of speech in the production of the vowels /i/, /a/ and /u/," J. Phonetics **17**, 123–133.

Rothenberg, M. (**1981**). "Acoustic interaction between the glottal source and the vocal tract," in *Vocal Fold Physiology*, edited by K. N. Stevens and M. Hirano (University of Tokyo Press, Tokyo), pp. 305–323.

Ruty, N., Pelorson, X., and Hirtum, A. V. (**2008**). "Influence of acoustic waveguides lengths on self-sustained oscillations: Theoretical prediction and experimental validation," J. Acoust. Soc. Am. **123**, 3121–3121.

Shadle, C. H. (**1985**). "Intrinsic fundamental frequency of vowels in sentence context," J. Acoust. Soc. Am. **78**, 1562–1567.

Sonderegger, M. A. (**2004**). "Subglottal coupling and vowel space: An investigation in quantal theory," Physics B.S. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Stevens, K. N. (**1971**). "Airflow and turbulence noise for fricative and stop consonants: Static considerations," J. Acoust. Soc. Am. **50**, 1180–1192.

Stevens, K. N. (**1977**). "Physics of laryngeal behavior and larynx modes," Phonetica **34**, 264–279.

Stevens, K. N., Kalikow, D. N., and Willemain, T. R. (**1975**). "A miniature accelerometer for detecting glottal waveforms and nasalization," J. Speech, Lang., Hear. Res. **18**, 594–599.

Titze, I., Riede, T., and Popolo, P. (**2008**). "Nonlinear source-filter coupling in phonation: Vocal exercises," J. Acoust. Soc. Am. **123**, 1902–1915.

Titze, I. R. (**1988**). "The physics of small-amplitude oscillation of the vocal folds," J. Acoust. Soc. Am. **83**, 1536–1552.

Titze, I. R. (**2008**). "Nonlinear source-filter coupling in phonation: Theory," J. Acoust. Soc. Am. **123**, 2733–2749.

Titze, I. R., and Story, B. H. (**1997**). "Acoustic interactions of the voice source with the lower vocal tract," J. Acoust. Soc. Am. **101**, 2234–2243.

Yegnanarayana, B., and Murty, K. S. R. (**2009**). "Event-based instantaneous fundamental frequency estimation from speech signals," IEEE Trans. Audio, Speech, Lang. Process. **17**, 614–624.

Zhang, Z., Neubauer, J., and Berry, D. A. (**2006**). "The influence of subglottal acoustics on laboratory models of phonation," J. Acoust. Soc. Am. **120**, 1558–1569.

Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 63–64 (2014) 70–83

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Extraction of formant bandwidths using properties of group delay functions

Anand Joseph Xavier Medabalimi [a,*], Guruprasad Seshadri [b], Yegnanarayana Bayya [a]

[a] International Institute of Information Technology, Hyderabad 500032, India
[b] Innovation Labs., TATA Consultancy Services, Bangalore 560066, India

Received 27 September 2013; received in revised form 6 March 2014; accepted 21 April 2014
Available online 9 May 2014

Abstract

Formant frequencies represent resonances of vocal tract system during the production of speech signals. Bandwidths associated with the formant frequencies are important parameters in analysis and synthesis of speech signals. In this paper, a method is proposed to extract the bandwidths associated with formant frequencies, by analysing short segments (2–3 ms) of speech signal. The method is based on two important properties of group delay function (GDF): (a) The GDF exhibits prominent peaks at resonant frequencies and (b) the influence of one resonant frequency on other resonances is negligible in GDF. The accuracy of the method is demonstrated for synthetic signals generated using all-pole filters. The method is evaluated by extracting bandwidths of synthetic signals in closed phase and open phase regions within a pitch period. The accuracy of the proposed method is also compared with that of two other methods, one based on linear prediction analysis of speech signals, and another based on filterbank arrays for obtaining amplitude envelopes and instantaneous frequency signals. Results indicate that the method based on the properties of GDF is suitable for accurate extraction of formant bandwidths, even from short segments of speech signal within a pitch period.

Keywords: Formant frequency; Bandwidth; Group delay function; Short segments; Closed phase; Open phase

1. Introduction

During the production of speech, the nature of speech sounds depends on the time-varying characteristics of the source of excitation and those of the vocal tract system. Formants are resonances of the vocal tract system, and they represent important sound-specific and speaker-specific information (Fant, 1960). Bandwidths associated with the formant frequencies are useful parameters in the analysis and synthesis of speech signals. Accurate extraction of formant bandwidths from speech signals is a difficult task, since the formant frequencies and their bandwidths vary across pitch periods. Formant frequencies and their bandwidths vary even within a pitch period, from closed phase of glottis to the open phase. This is due to decoupling of trachea and vocal tract during the closed phase of glottis, and coupling of source-tract during the open phase of glottis. These issues necessitate analysis of short segments of speech (typically less than a pitch period) for extraction of formant bandwidths.

Some methods proposed in the literature for the extraction of formant bandwidths use model-based approaches for representation of speech signal. An approach based on AM–FM modeling of speech signal was proposed by Cohen et al. (1992), and an expression for formant bandwidth was obtained in terms of the parameters of the model. Potamianos and Maragos (1995) employed a

* Corresponding author.
E-mail addresses: anandjm@research.iiit.ac.in (A.J.X. Medabalimi), guruprasad.seshadri@gmail.com (G. Seshadri), yegna@iiit.ac.in (Y. Bayya).

http://dx.doi.org/10.1016/j.specom.2014.04.006
0167-6393/

bank of Gabor bandpass filters to decompose the speech signal, and the signal in each band was demodulated to obtain amplitude envelope and instantaneous frequency signals. Bandwidth estimates were obtained from the instantaneous frequency signals. An exponentially weighted autoregressive (EWAR) spectral model was proposed to extract the bandwidths of formant frequencies (Zheng and Hasegawa-Johnson, 2003). A method called clustered line-spectrum modeling was proposed to decompose the speech signal into three dominant resonant oscillations with nearly exponentially decaying envelopes (Yasojima et al., 2006). The bandwidths were estimated from the decaying constants of the resonant frequencies. In the above cases, the accuracy of extraction of bandwidths depends on the fitness of the models, and on the accuracy of estimation of model parameters from the speech signal. These methods typically use more than one pitch cycle of speech signal, a duration over which the bandwidths of formants tend to vary. Hence, there is need for methods to extract bandwidths from short segments (2–3 ms in duration) of voiced speech signals.

Linear prediction (LP) analysis is commonly used for extracting formant frequencies (Makhoul, 1975). The estimation of autoregressive parameters in LP analysis is based on an error minimization criterion. Reddy and Swamy (1984) proposed a method to extract bandwidths of formants, by observing the phase slope of the z-transform around the poles obtained using LP analysis of speech signal. However, the accuracy of formant frequencies and bandwidths depends on the choice of order of LP analysis. Also, the error minimization criterion in LP analysis focuses on matching spectral peaks, and bandwidth is only an additional outcome of the process.

In this paper, we propose a method for extraction of formant bandwidths by exploiting the properties of phase of Fourier transform. The method assumes that the speech signal in voiced regions can be modeled as the output of an all-pole filter. The key idea is based on the evaluation of group delay function at the resonant frequencies. In Section 2, some important properties of phase response of all-pole systems are revisited. These properties are exploited in Section 3, which describes the analytical basis for the proposed method. This section also examines the effectiveness of bandwidth extraction for all-pole systems. The method is evaluated for the case of synthetic speech signals in Section 4. Accuracy of the method is compared with two other methods of bandwidth extraction, and results are discussed. Conclusions are given in Section 5.

2. Properties of phase response of all-pole systems

In this section, we summarize the observations reported by Yegnanarayana (1978) on the significance of processing phase response and its derivative for extraction of formant frequencies from speech signals. These observations provide a background for the method proposed in Section 3.1, for extraction of formant bandwidths from discrete-time speech signals.

Let us consider a cascade of M resonators. The frequency response of the i^{th} resonator is given by

$$H_i(\omega) = \frac{1}{(j\omega - (\alpha_i + j\beta_i))(j\omega - (\alpha_i - j\beta_i))}, \tag{1}$$

where $\alpha_i \pm j\beta_i$ is the complex pair of poles of the i^{th} resonator, ω is the analog angular frequency and $j = \sqrt{-1}$. The expression is simplified as

$$H_i(\omega) = \frac{1}{\alpha_i^2 + \beta_i^2 - \omega^2 - 2j\omega\alpha_i}. \tag{2}$$

The squared magnitude response of the i^{th} resonator is given by

$$|H_i(\omega)|^2 = \frac{1}{(\alpha_i^2 + \beta_i^2 - \omega^2)^2 + 4\omega^2\alpha_i^2}. \tag{3}$$

The squared magnitude response of the overall filter, i.e., the cascade of M resonators, is given by

$$|H(\omega)|^2 = \prod_{i=1}^{M} |H_i(\omega)|^2. \tag{4}$$

The phase response of the i^{th} resonator is given by

$$\Theta_i(\omega) = \tan^{-1}\left(\frac{2\omega\alpha_i}{\alpha_i^2 + \beta_i^2 - \omega^2}\right). \tag{5}$$

Group delay function of the i^{th} resonator, which is the negative derivative of the corresponding phase response, is given by Yegnanarayana (1978)

$$G_i(\omega) = -\frac{2\alpha_i(\alpha_i^2 + \beta_i^2 + \omega^2)}{(\alpha_i^2 + \beta_i^2 - \omega^2)^2 + 4\omega^2\alpha_i^2}. \tag{6}$$

The group delay function of the overall filter is given by

$$G(\omega) = \sum_{i=1}^{M} G_i(\omega). \tag{7}$$

It can be verified that the magnitude response $|H_i(\omega)|^2$ (Eq. (3)) has a peak at $\omega^2 = \beta_i^2 - \alpha_i^2$, and a half power bandwidth of α_i. For sharp resonant peaks in the magnitude response, $\beta_i^2 \gg \alpha_i^2$. We now note the following properties of $G_i(\omega)$ (Yegnanarayana, 1978).

(a) From Eq. (3) and (6)

$$G_i(\omega) = -2\alpha_i(\alpha_i^2 + \beta_i^2 + \omega^2)|H_i(\omega)|^2. \tag{8}$$

For $\beta_i^2 \gg \alpha_i^2$, the group delay function $G_i(\omega)$ can be approximated around the resonant frequency $\omega^2 = \beta_i^2 - \alpha_i^2$ as follows:

$$G_i(\omega) = K_i|H_i(\omega)|^2, \tag{9}$$

where K_i is a constant. It is to be noted that $G_i(\omega)$ too has a peak near $\omega^2 = \beta_i^2 - \alpha_i^2$.

(b) At frequencies much lower than the resonant frequency $\beta_i^2 - \alpha_i^2$, $G_i(\omega) \approx -\frac{2\alpha_i}{\beta_i^2}$, which is a small constant quantity (since $\beta_i^2 \gg \alpha_i^2$).
(c) At frequencies much higher than the resonant frequency $\beta_i^2 - \alpha_i^2$, $G_i(\omega) \approx -\frac{2\alpha_i}{\omega^2}$.

From (a) above, we observe that resolution of neighboring formants with different bandwidths is easier from the group delay function than from the magnitude response, due to two reasons: (i) The group delay function is proportional to the square of magnitude response in the neighborhood of resonant frequencies (Eq. (9)). (ii) The group delay function corresponding to the overall filter is the summation of the individual group delay functions (Eq. (7)), whereas the magnitude response corresponding to the overall filter is the product of the individual magnitude responses (Eq. (4)). From the properties (b) and (c) above, we observe that the group delay function of a given resonator has relatively less influence on that of the other resonators, since the group delay function assumes low values for frequencies that are much smaller or much larger that the resonant frequency. Although the above observations are made in the context of analog signals, we verify that the observations are valid in the context of discrete-time signals too.

3. Basis for the proposed method

We first establish a relation between the bandwidth of a resonant frequency and the group delay function of the corresponding all-pole system. The bandwidth is then expressed in terms of the group delay function computed from the discrete-time signal. Extraction of bandwidth is discussed for three cases, namely, all-pole systems described by (a) a single complex pole, (b) a pair of complex conjugate poles, and (c) multiple pairs of complex conjugate poles. Some issues in the computation of group delay function from natural speech signals are also discussed in this section.

3.1. Relation between bandwidth and group delay function for discrete-time signals

Let us consider a discrete-time signal, which can be modeled as the output of a single-pole system. Let $H(z) = \frac{1}{1-z_0 z^{-1}}$ represent the transfer function of a system which has a pole at $z = z_0$, where $z_0 = r_0 e^{j\omega_0}$. Here, r_0 denotes the radial distance of the pole from the origin ($z = 0$) and ω_0 denotes the resonant frequency in radians. The frequency response of the system is given by Oppenheim et al. (1999)

$$H(\omega) = \frac{1}{1 - r_0 e^{j\omega_0} e^{-j\omega}}. \tag{10}$$

In this case, the impulse response is complex. The frequency response consists of magnitude response and phase response. Group delay function (GDF) is defined as the negative derivative of the phase of Fourier transform. For the system described by Eq. (10), the GDF is given by Oppenheim et al. (1999)

$$\tau(\omega) = -\frac{r_0^2 - r_0 \cos(\omega - \omega_0)}{1 + r_0^2 - 2r_0 \cos(\omega - \omega_0)}. \tag{11}$$

The GDF can also be computed from the samples of a given discrete-time signal $x[n]$. Let $X(\omega) = \sum_{n=-\infty}^{\infty} x[n] e^{-j\omega n}$ denote the discrete-time Fourier transform of $x[n]$. Also, $X(\omega)$ can be expressed in terms of its magnitude $|X(\omega)|$ and phase $\phi(\omega)$ as $X(\omega) = |X(\omega)| e^{j\phi(\omega)}$. Applying logarithmic transformation on both sides, and differentiating with respect to ω, we have (Oppenheim and Schafer, 1975a)

$$\frac{1}{X(\omega)} \frac{d}{d\omega}(X(\omega)) = \frac{d}{d\omega}(\log |X(\omega)|) + j \frac{d}{d(\omega)}(\phi(\omega)). \tag{12}$$

Since GDF is defined as the negative derivative of phase of Fourier transform, it follows (from Eq. (12)) that $g(\omega)$ can be expressed as

$$g(\omega) = -\mathrm{Im}\left(\frac{1}{X(\omega)} \frac{d}{d\omega}(X(\omega)) \right). \tag{13}$$

We make use of the relation that the Fourier transform of $y[n] = nx[n]$ is given by $Y(\omega) = j\frac{d}{d\omega}(X(\omega))$ (Oppenheim and Schafer, 1975b). Thus,

$$g(\omega) = -\mathrm{Im}\left(\frac{1}{j} \frac{Y(\omega)}{X(\omega)} \right). \tag{14}$$

The above equation can be simplified as (Oppenheim and Schafer, 1975a)

$$g(\omega) = \frac{X_R(\omega) Y_R(\omega) + X_I(\omega) Y_I(\omega)}{X_R^2(\omega) + X_I^2(\omega)}, \tag{15}$$

where $X(\omega) = X_R(\omega) + jX_I(\omega)$ and $Y(\omega) = Y_R(\omega) + jY_I(\omega)$, and the subscripts R and I denote the real and imaginary parts, respectively.

Thus, $\tau(\omega)$ (in Eq. (11)) gives the analytical expression for the GDF of a digital resonator represented by a single pole, while $g(\omega)$ (in Eq. (15)) is the GDF as computed from the discrete-time signal $x[n]$. If $x[n]$ is assumed to be the impulse response of a single-pole digital resonator, then the two representations of GDF are equivalent. This equivalence can be exploited to obtain bandwidths associated with resonant frequencies.

Assuming $x[n]$ to be the impulse response of a single-pole digital resonator, the GDF computed from segments of $x[n]$ using Eq. (15) should be equivalent to the GDF obtained using Eq. (11), for all values of ω. However, the GDF computed from a discrete-time signal is affected by the length, shape and location of the analysis window. Fig. 1(a) shows the GDF of an all-pole filter with a single pole (as obtained by Eq. (11)), which has a resonant frequency at 2,000 Hz and $r_0 = 0.939$ (corresponding to a bandwidth of 200 Hz at a sampling frequency of 10,000 Hz). Fig. 1(b) shows the GDF computed from the samples of the impulse response of the all-pole filter

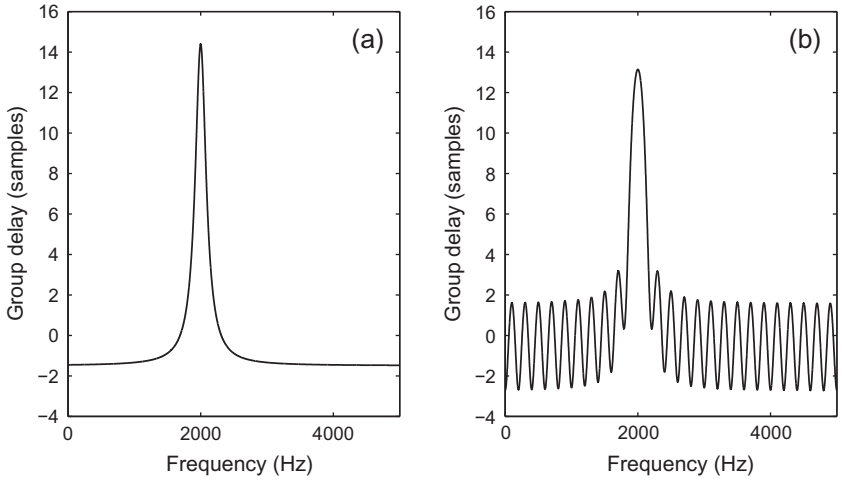


Fig. 1. (a) Group delay function (GDF) of a single pole digital resonator. (b) GDF computed from the samples of impulse response of the resonator.

(as obtained by Eq. (15)). For generating the impulse response, sampling frequency of 10000 Hz was used. The GDF was computed from a 5 ms window of the impulse response. The effect of windowing is visible in Fig. 1(b). The correspondence between the two representations of GDF is observed to be most prominent at the resonant frequency (2000 Hz). We evaluate the two representations of GDF at $\omega = \omega_0$, and equate them. From Eq. (11), $\tau(\omega_0) = \frac{r_0}{1-r_0}$. From Eqs. (11) and (15), we have

$$r_0 = \frac{g(\omega_0)}{1 + g(\omega_0)}, \tag{16}$$

where $g(\omega_0)$ is evaluated using Eq. (15). Thus, r_0 associated with a single pole can be expressed in terms of its GDF. The bandwidth β (in Hz) of the resonator can be computed from the radial distance r_0 of the pole using the relation $r_0 \approx e^{-\pi\beta T}$, where T (in seconds) is the sampling interval of the discrete-time signal. We will discuss the significance of length of analysis window in Section 3.4.

We now examine the accuracy of extraction of bandwidth for three cases, where the all-pole filter is defined by (a) a single complex pole, (b) a pair of complex conjugate poles, and (c) multiple pairs of complex conjugate poles. In each case, the values of resonant frequency and its bandwidth are assumed in order to generate the impulse response of the all-pole filter. The GDF is then computed from the samples of the impulse response, and is compared to the analytical expression of GDF for that all-pole filter. This comparison results in a solution for the bandwidth associated with the resonant frequency.

3.2. Single complex pole

The radial distance r_0 of the single complex pole can be computed from Eq. (16), as discussed above. We denote the value computed from Eq. (16) as $\hat{r}_0$, to distinguish it from the true/reference value r_0. The accuracy of extraction of r_0 is evaluated by generating the impulse response of the resonator for different values of r_0 and ω_0. The value of r_0 is varied between 0.90 and 0.995, in steps of 0.01. The value of $f_0 = \frac{\omega_0}{2\pi}$ is varied between 300 Hz and 4,000 Hz, in steps of 50 Hz. A sampling frequency of 10,000 Hz was used. The percentage error in the extraction of r_0 is given by $\delta = \frac{\hat{r}_0 - r_0}{r_0} \times 100$. The GDF is computed from the impulse response using four different window lengths (W), namely (a) 2 ms, (b) 3 ms, (c) 5 ms and (d) 10 ms. The distribution of error for the four cases is shown in Fig. 2. The average (E) of the absolute values of error is also indicated in the figure. Due to the effect of windowing, the estimated radial distance $\hat{r}_0$ of the pole is lesser than the true value r_0. Hence, the distributions in Fig. 2 are skewed toward negative values of δ. The method estimates a higher bandwidth (in Hz) relative to the true value. Smaller the window length, larger is the estimated bandwidth and greater is the spread of δ. The spread of the error reduces as the length W of the analysis window increases, since the accuracy of computation of GDF improves with increase in W. Computation of GDF using Eq. (15) is affected by the sampling rate. A higher sampling rate is desirable for accurate estimation of bandwidth, particularly for short windows of analysis. The effect of windowing is more pronounced for short windows of analysis ($W = 2$ ms and $W = 3$ ms), than for larger windows ($W = 5$ ms and $W = 10$ ms).

3.3. Single pair of complex conjugate poles

Let us consider an all-pole filter defined by a pair of complex conjugate poles. The transfer function of the digital resonator in this case is given by $H(z) = \frac{1}{(1-z_0 z^{-1})(1-z_0^* z^{-1})}$, where $z_0 = r_0 e^{j\omega_0}$ and z_0^* denotes the complex conjugate of z_0. The impulse response of the digital resonator is a real signal. Due to additive nature of phase, the GDF has two terms, corresponding to the poles at $\omega = \omega_0$ and $\omega = -\omega_0$. The GDF is given by $\tau(\omega) = \tau_+(\omega) + \tau_-(\omega)$, where

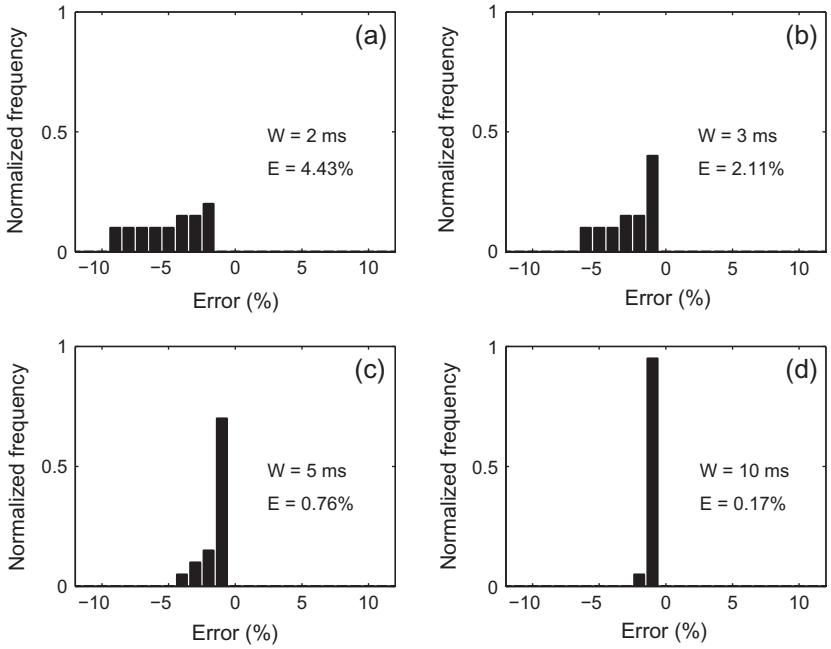


Fig. 2. Case of an all-pole filter which is represented by a single complex pole. Distribution of error in the extraction of bandwidth is shown for different lengths (W) of the analysis window, namely, (a) $W = 2$ ms, (b) $W = 3$ ms, (c) $W = 5$ ms, and (d) $W = 10$ ms. The average (E) of the absolute values of error is also shown.

$$\tau_+(\omega) = -\frac{r_0^2 - r_0\cos(\omega - \omega_0)}{1 + r_0^2 - 2r_0\cos(\omega - \omega_0)},$$
$$\tau_-(\omega) = -\frac{r_0^2 - r_0\cos(\omega + \omega_0)}{1 + r_0^2 - 2r_0\cos(\omega + \omega_0)}. \quad (17)$$

In this case, the peaks in $\tau(\omega)$ do not occur exactly at $\omega = \pm\omega_0$, but are shifted depending on the value of the bandwidth (Yegnanarayana, 1978). The shift is smaller for poles closer to the unit circle ($|z| = 1$). The value of r_0 can be obtained by solving $\tau(\omega_0) = g(\omega_0)$, where $g(\omega_0)$ and $\tau(\omega_0)$ are evaluated using Eqs. (15) and (17), respectively. The exact solution involves a third order polynomial in r_0. An approximate solution can be obtained by observing the condition under which the effect of $\tau_-(\omega)$ on $\tau(\omega)$ at $\omega = \omega_0$ can be ignored. The contribution of $\tau_-(\omega)$ to $\tau(\omega)$ at $\omega = \omega_0$ can be measured in terms of the ratio α, which is given by $\alpha = \frac{|\tau_+(\omega_0)|}{|\tau_-(\omega_0)|}$. To observe the variation of α, the value of r_0 is varied between 0.90 and 0.995, in steps of 0.01. The value of $f_0 = \frac{\omega_0}{2\pi}$ is varied between 300 Hz and 4,000 Hz, in steps of 50 Hz. A sampling frequency of 10,000 Hz was used. In all the cases, the value of α was observed to be greater than 19, indicating that the contribution of $\tau_-(\omega_0)$ to $\tau(\omega_0)$ can be ignored. Thus, $\tau(\omega_0) \approx \frac{r_0}{1-r_0}$, and the solution to obtain r_0 in the case of a pair of complex conjugate poles is same as that in the case of a single complex pole. However, for smaller values of r_0, the contribution of $\tau_-(\omega_0)$ to $\tau(\omega_0)$ is significant. Fig. 3 shows the accuracy of extraction of r_0 for the case of a pair of complex conjugate poles. The GDF is computed from the impulse response of the digital resonator using four different window lengths (W), namely (a) 2 ms, (b) 3 ms, (c) 5 ms and (d) 10 ms. The average (E) of the absolute values of error is also indicated in the figure. The accuracy in this case is similar to that of the single complex pole, indicating that the effect of $\tau_-(\omega_0)$ on $\tau(\omega_0)$ is negligible for larger values of r_0.

3.4. Multiple pairs of complex conjugate poles

Let us consider an all-pole system defined by N pole pairs. Let the poles of the system be represented by z_k and z_k^*, where $z_k = r_k e^{j\omega_k}, k = 1, \ldots, N$. Here, r_k and ω_k represent the radial distance and resonant frequency, respectively, of the k^{th} pole. The GDF is given by

$$\tau(\omega) = \sum_{k=1}^{N}\{\tau_{k,+}(\omega) + \tau_{k,-}(\omega)\}, \text{ where}$$
$$\tau_{k,+}(\omega) = -\frac{r_k^2 - r_k\cos(\omega - \omega_k)}{1 + r_k^2 - 2r_k\cos(\omega - \omega_k)}, \quad (18)$$
$$\tau_{k,-}(\omega) = -\frac{r_k^2 - r_k\cos(\omega + \omega_k)}{1 + r_k^2 - 2r_k\cos(\omega + \omega_k)}.$$

In the case of multiple resonances, the influence of one resonant peak on the other peaks is negligible in group delay function, if the formants have relatively small bandwidths (Yegnanarayana, 1978). This observation is valid even for closely spaced formants, but it is not valid for resonant frequencies with large bandwidths (Yegnanarayana, 1978). When the GDF is evaluated at $\omega = \omega_i$, the influence of the additional $2N - 1$ terms can be measured in terms of the ratio $\alpha_i, i = 1, \ldots, N$, given by

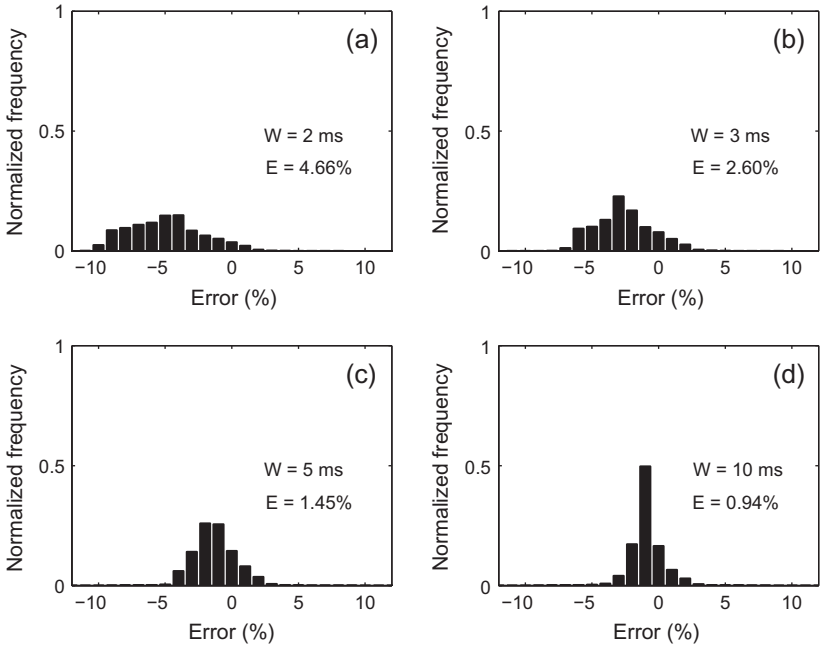


Fig. 3. Case of an all-pole filter which is represented by a pair of complex conjugate poles. Distribution of error in the extraction of bandwidth is shown for different lengths (W) of the analysis window, namely, (a) $W = 2$ ms, (b) $W = 3$ ms, (c) $W = 5$ ms, and (d) $W = 10$ ms. The average (E) of the absolute values of error is also shown.

$$\alpha_i = \frac{|\tau_{i,+}(\omega_i)|}{|\tau_{i,-}(\omega_i) + \sum_{k\neq i}\{\tau_{k,+}(\omega_i) + \tau_{k,-}(\omega_i)\}|}. \quad (19)$$

Here, the denominator represents the contribution of other resonances on the GDF evaluated at $\omega = \omega_i$. We consider the case of three pole pairs, i.e., $N = 3$, for computing the values of α_i. This computation requires the choice of three resonant frequencies and the corresponding bandwidths. The values of resonant frequencies are chosen from vocal tract resonance (VTR) database (Deng et al., 2006). This choice is relevant in the context of analysis of voiced speech signals, which typically consist of three or four prominent resonances. The formant frequencies of speech signals in the database have been marked in a semi-automatic manner, by observing speech signals and their spectrograms. Another reason for using formant frequencies from VTR database is to simulate the time varying nature of formants in voiced speech signals.

Formant frequencies marked from 50 speech utterances of VTR database are used, leading to 14,620 sets of values, where each set consists of first three formant frequencies (denoted by F_1, F_2 and F_3). Two cases are considered: (i) Bandwidths are assumed to be same for all the three formants, corresponding to $r_k = 0.97, k = 1, 2, 3$. (ii) Bandwidths are assumed to be one-tenth (10%) of the formant frequencies. In cases (i) and (ii), the values of α_1, α_2 and α_3 are computed. Note that this computation is done according to Eq. (19), and does not involve analysis of signals. For case (i), distributions of α_1, α_2 and α_3 are shown in Fig. 4(a)–(c), respectively. The mean values μ_1, μ_2 and μ_3 of the distributions indicate that the effect of the other two formants on a given formant is less than 7%. For case (ii), distributions of α_1, α_2 and α_3 are shown in Fig. 4(d)–(f), respectively. The values of μ_1 and μ_2 indicate that the effect of the other two formants on a given formant is less than 6% in the case of F_1 and F_2. Contribution of the term $\tau_{3,-}(\omega_3)$ to $\tau(\omega_3)$ is significant (about 16%) due to larger bandwidth of F_3. This results in lower value of μ_3, compared to those of μ_1 and μ_2. These observations give a basis for ignoring the effect of other resonances on the GDF of the resonance of interest, when computing bandwidths.

In the context of speech signals, bandwidths are related to formant frequencies. Hence, we consider case (ii) for extraction of bandwidths. The impulse response of all-pole filter is generated at sampling frequency of 10,000 Hz using the formant frequencies from VTR database, and by assuming the bandwidths to be one-tenth (10%) of the formant frequencies. The impulse response is analysed and formant bandwidths are extracted using the relation $r_i = \frac{g(\omega_i)}{1+g(\omega_i)}, i = 1, 2, 3$. The GDF is computed from the impulse response for two different window lengths ($W = 3$ ms and $W = 10$ ms). Distributions of errors in the extraction of bandwidths corresponding to the three formant frequencies are shown in Fig. 4(g)–(i) for $W = 3$ ms, and in Fig. 4(j)–(l) for $W = 10$ ms. The averages (E_i) of the absolute values of errors are also indicated in Fig. 4(g)–(l). Errors in the extraction of bandwidths are lower when a longer analysis window ($W = 10$ ms) is used, compared to the errors when a shorter analysis window ($W = 3$ ms) is used. The low error rates (even for short analysis window) indicate the accuracy of the proposed method.

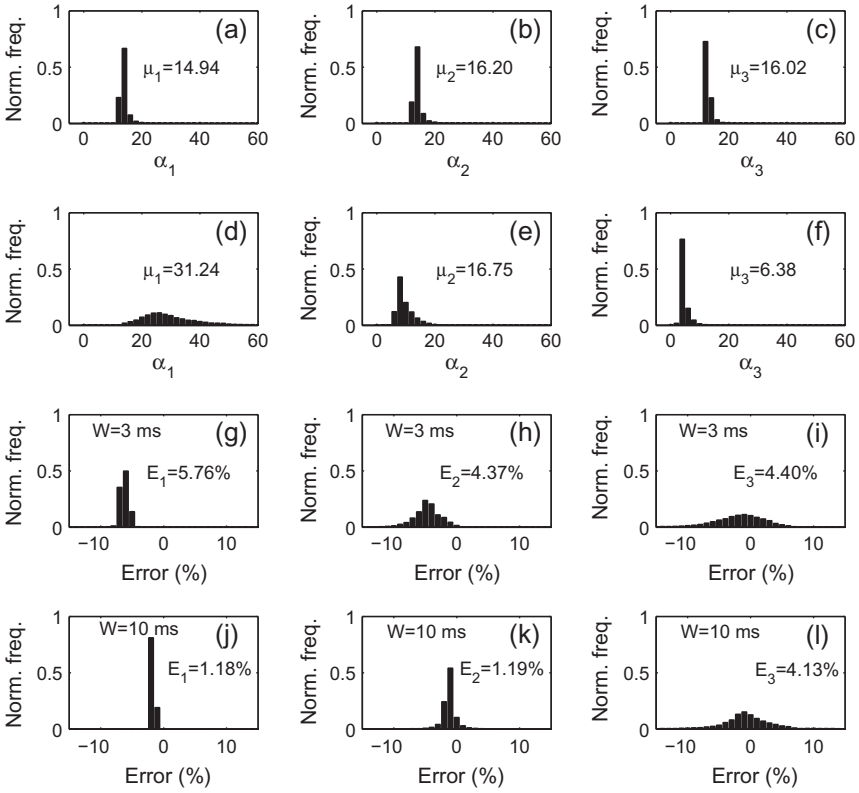


Fig. 4. Case of an all-pole filter which is represented by three pairs of complex conjugate poles. (a), (b) and (c) Show the distribution of α_1, α_2 and α_3, respectively, when equal bandwidths are assumed. (d), (e) and (f) Show the distribution of α_1, α_2 and α_3, respectively, when bandwidths are assumed to be one-tenth of the corresponding formant frequencies. For the case of unequal bandwidths, distribution of error in bandwidth extraction is shown for the first three formants in (g), (h) and (i), respectively, for $W = 3$ ms, and in (j), (k) and (l), respectively, for $W = 10$ ms.

3.5. Computation of group delay function from natural speech signals

In voiced regions, speech signal can be modeled as the output of an all-pole filter which is defined by multiple pairs of complex conjugate poles. The signal segment within a pitch period can be modeled as the impulse response of the all-pole filter. However, the excitation component of speech signal is not an ideal impulse. Also, the vocal tract system is not exactly equivalent to an all-pole system. Due to these factors, the GDF computed from natural speech signals differs somewhat from the GDF computed from synthetic signals. Fig. 5 illustrates the difference between the group delay functions computed from natural and synthetic speech signals. A segment of natural speech signal corresponding to one pitch period (of 10 ms) is shown in Fig. 5(a). This is a segment between two successive instants of glottal closure. A segment of synthetic speech signal, corresponding to the natural speech signal, is shown in Fig. 5(c) and (e). The synthetic speech signal is obtained as follows:

(a) The segment of natural speech (shown in Fig. 5(a)) is analyzed to extract the first three formant frequencies.
(b) For each formant frequency, a 3 dB bandwidth equal to 10% of the formant frequency is assumed.
(c) An all-pole filter is constructed by using the formant frequencies and their corresponding bandwidths.
(d) The all-pole filter is excited using a unit sample sequence.

The signals in Fig. 5(a) and (b) are sampled at 32,000 Hz. Fig. 5(b) and (d) shows the GDF computed from the segments of natural and synthetic speech signals, respectively. In both the cases, the GDFs are computed using Eq. (15). In the expression for GDF computed from signal samples (Eq. (15)), the denominator is the square of magnitude of the short-time spectrum of the signal. Due to the effect of windowing, the term in the denominator has values that are close to zero, causing the GDF to assume very large values. Such large values in the GDF can be observed in both Fig. 5(b) and (d). Peaks due to all the three formant frequencies are prominent in Fig. 5(d). This is due to two reasons:

(i) In the neighborhood of resonant frequencies, the GDF is proportional to the square of magnitude of the short-time spectrum.
(ii) The signal is the output of an all-pole filter.

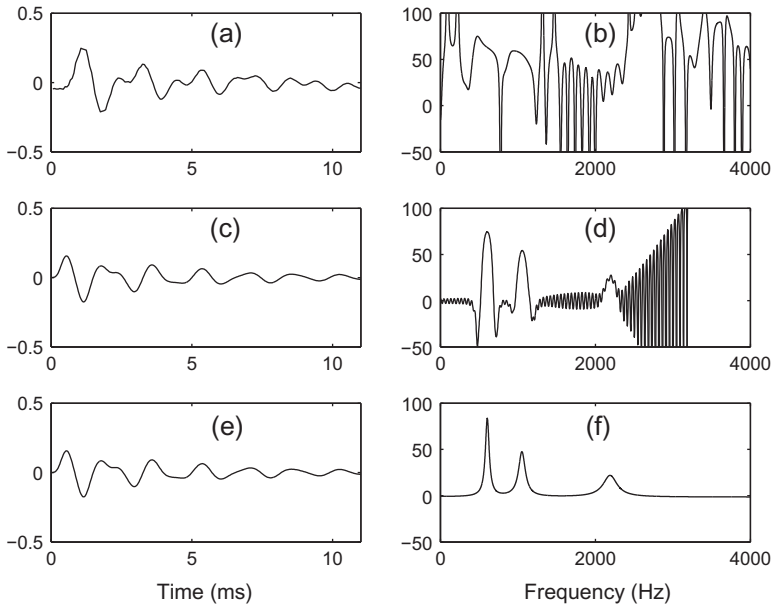


Fig. 5. (a) A segment of natural speech signal and (b) GDF computed from the segment. (c) A segment of synthetic speech signal (synthesized using the first three formant frequencies of the natural speech segment). (d) GDF computed from the signal in (c). (e) This signal is same as the signal shown in (c). (f) GDF of the all-pole filter used for synthesizing the signal in (e).

Peaks due to the first two formant frequencies are visible in Fig. 5(b) also, due to the reason (i) given above. However, the signal Fig. 5(a) is not exactly the output of an all-pole model, and hence, the GDF in Fig. 5(b) appears different from that in Fig. 5(d). The third formant frequency is not prominent at all in Fig. 5(b). Fig. 5(f) shows the GDF of the all-pole filter, as given by Eq. (18). All the three formant frequencies are prominent in Fig. 5(f). It is to be noted that the $\tau(\omega)$ in Eq. (18) is effectively derived from the infinite length impulse response of the all-pole filter, whereas $g(\omega)$ (in Eq. (15)) is computed from a finite length of the signal. Hence, the GDF in Fig. 5(f) is free of the effect of windowing. Also, the values of GDF at the formant frequencies are similar (but not same) across Fig. 5(b), (d), and (f). When computing $g(\omega)$ (as given by Eq. (15)) from samples of synthetic or natural speech signal, the length of analysis window should be less than or equal to a pitch period. Such a choice ensures that $g(\omega)$ is closer to $\tau(\omega)$, where the latter is computed using Eq. (18). If $g(\omega)$ is computed from more than one pitch period of speech signal, the equivalence between $g(\omega)$ and $\tau(\omega)$ is not valid. Hence, it is necessary to ensure that the analysis window is less than or equal to one pitch period, and that the analysis window begins immediately after an instant of glottal closure.

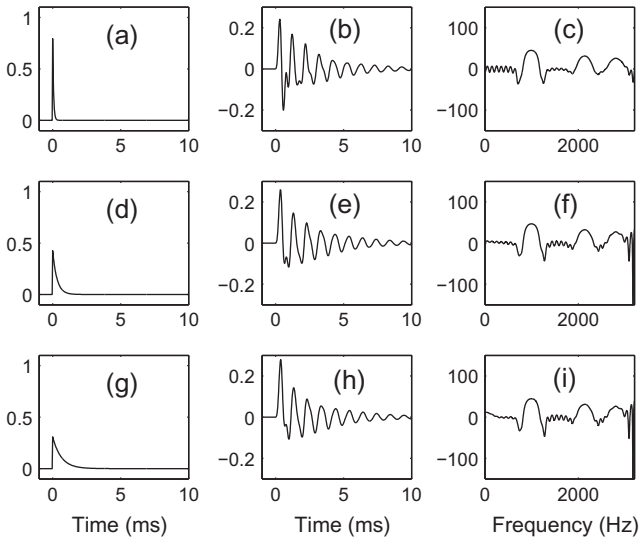


Fig. 6. Plots (a), (d) and (g) show three different impulse-like excitation signals, with varying sharpness/abruptness. All the excitation signals have the same value of overall energy. Plots (b), (e) and (h) show the signals synthesized by exciting an all-pole filter with input signals shown in (a), (d) and (g), respectively. Plots (c), (f) and (i) show the group delay functions (in samples) computed from the signals shown in (b), (e) and (h), respectively.

We now examine the effect of the nature of excitation on the group delay function. Fig. 6(a), (d) and (g) represent three different impulse-like excitation signals, but with varying sharpness/abruptness. However, all the three signals have the same value of overall energy. The signal in Fig. 6(a) is closest to an ideal impulse (unit sample sequence), while the signals in Fig. 6(d) and (g) progressively deviate from the ideal impulse. Fig. 6(b), (e) and (h) shows the signals synthesized using the excitation signals in Fig. 6(a), (d) and (g), respectively. The formant frequencies and bandwidths used for synthesis were the same for each excitation signal. A pitch period of 10 ms and a sampling frequency of 32,000 Hz were used for synthesis. The GDFs computed from the synthesized signals are shown in Fig. 6(c), (f) and (i). The effect of change in the impulse-like nature of excitation signal on the characteristics of GDF is not significant. The peaks due to formant frequencies are prominent in Fig. 6(c), (f) and (i)

with similar values of GDF at formant frequencies. It has been observed that a change in the impulse-like nature of the excitation signal affects the spectral tilt/slope of the synthesized signal, but not the position of formant frequencies (Gauffin and Sundberg, 1989). It is the response of vocal tract system which has a greater influence on the nature of GDF, than the excitation signal. This is mainly due to the presence of zeros in the short-time magnitude spectrum of the signal, which affect the nature of GDF.

4. Evaluation of bandwidth extraction methods for synthetic speech signals

In this section, we describe the synthesis of speech signals with known values of formant frequencies and their bandwidths. The synthesized signals are used to evaluate three different methods of extraction of bandwidths. These methods are described in the section. Accuracy of bandwidths extracted from the three methods are compared.

4.1. Synthesis of signals for evaluation

Extraction of formant bandwidths from speech signals is a difficult problem, primarily due to time-varying characteristics of speech signal. Formant frequencies vary across pitch periods, and even within a single pitch period, during the open and closed phases of glottis. During the closed phase of glottis, there is no coupling (or minimal coupling) between the trachea and the vocal tract. During the open phase of glottis, the trachea and the vocal tract are coupled, resulting in an increase in the effective length of the vocal tract. Hence, formant frequencies have slightly lesser values during the open phase, compared to those during the closed phase. Also, there is greater damping of resonances during the open phase, compared to that during the closed phase. Hence, formant bandwidths are greater during the open phase compared to the closed phase.

In order to evaluate the accuracy of a given method of extraction of formant bandwidths from speech signals, the true/reference values of formant frequencies and their bandwidths are required. While manual/semi-automatic labeling of formant frequencies from speech signals and their spectrograms is reliable, the same is not true for bandwidths. Hence, we generate synthetic signals using known formant frequencies, and by assuming 3 dB bandwidths as fractions of the corresponding formant frequencies. We use VTR database, where formant frequencies are marked corresponding to speech signals for 516 utterances. The formant frequencies are marked for every 10 ms. The synthesis of speech signals is performed as follows:

(a) In the closed phase of glottis, coefficients of the all-pole filter are derived using three formant frequencies (F_1, F_2 and F_3) and their bandwidths ($F_1/10, F_2/10$ and $F_3/10$). In the open phase of glottis, coupling between vocal source and vocal tract results in an increase in the effective length of the vocal tract. This causes a slight decrease (of about 3–6%) in the values of formant frequencies in the open phase relative to those in the closed phase. Also, damping of resonances in the open phase is greater than that in the closed phase. Hence, in the open phase of glottis, coefficients of the all-pole filter are derived using three formant frequencies whose values are $0.95F_1, 0.95F_2$ and $0.95F_3$, and the corresponding bandwidths whose values are $0.95F_1/8, 0.95F_2/8$ and $0.95F_3/8$. Thus, the coefficients of the all-pole filter change within each pitch period.

(b) In each pitch period, first half of the pitch period is considered as the closed phase, while the second half is considered as the open phase. In a pitch period consisting of $2L$ samples (where L is an integer), the glottal closure instant is located at the first sample while the glottal opening instant is located at $(L+1)^{\text{th}}$ sample.

(c) The amplitude of excitation signal is 1 unit at the instant of glottal closure, and 0.1 unit at the instant of glottal opening. The excitation signal consists of zeros at all other sample indices within each pitch period.

(d) For each utterance in the VTR database, three signals are synthesized, corresponding to pitch periods of 5 ms, 8 ms and 10 ms. Although the formant frequencies are marked for every 10 ms in VTR database, we use one set of formant frequencies (F_1, F_2 and F_3) to synthesize the signal within one pitch period. The pitch period in a synthesized signal is constant, but the formant frequencies vary from one pitch period to another.

(e) All signals are synthesized at two sampling rates, 10,000 Hz and 16,000 Hz.

4.2. Methods for extraction of bandwidths

We evaluate three methods of extraction of formant bandwidths from the synthesized speech signals. These are: (a) The proposed method based on the properties of group delay functions (GDF), (b) the method based on linear prediction analysis (LPA) (Makhoul, 1975), and (c) the method proposed by Potamianos and Maragos (1995), which processes speech signal through a filter bank array (FBA). In each method, we use the true locations of glottal closure instants and glottal opening instants, so that the regions of closed phase and open phase are available for analysis. Signal segments in the regions of closed phase and open phase are analyzed for extraction of formant bandwidths. The three methods for extraction of formant bandwidths are summarized below.

4.2.1. Method based on properties of GDF

This method has been described in Section 3.4. In this method, the knowledge of formant frequencies is necessary to extract the bandwidths. Formant frequencies are

extracted from the regions of open and closed phase of glottis, using the method proposed by Xavier et al. (2006). The method proposed by Xavier et al. (2006) for extraction of formant frequencies is also based on the properties of GDF. The GDF is computed from the samples of the speech signal in the regions of closed/open phase. Using the knowledge of formant frequencies, radial distances corresponding to the formant frequencies are derived using the proposed method (Section 3.4). For extraction of formants and bandwidths, signals synthesized at 16,000 Hz are used.

Since the proposed method depends on the knowledge of formant frequencies for extraction of bandwidths, it is necessary to evaluate the accuracy of the method used for extraction of formants. Let F denote the true value of a formant frequency, and let $\widehat{F}$ denote the value extracted from synthetic signal. The percentage deviation δ_F is given by $\delta_F = \frac{|F-\widehat{F}|}{F} \times 100$. Table 1 shows the accuracy of formant extraction, when the formant frequencies are extracted using the method proposed by Xavier et al. (2006). In the table, each entry denotes the percentage of pitch periods for which the percentage deviation δ_F is lesser than a certain value (such as 2%, 5% and 10%). It is observed from Table 1 that the method of formant extraction is more accurate in the closed phase than in the open phase. In more than 90% of pitch periods, formant frequencies are extracted with a percentage deviation of less than 10%. Also, the accuracy of formant extraction improves as the pitch period increases from 5 ms to 10 ms, since a greater length of signal segment is available for analysis at longer pitch periods. The effect of accuracy of formant extraction on that of bandwidth extraction is discussed in Section 4.3.

Fig. 7 illustrates the extraction of formant bandwidths using the proposed method, from short segments (3 ms) of speech signal in the regions of closed and open phases of glottis. Glottal closure instants are detected using the method proposed by Murty and Yegnanarayana (2008). This method has high accuracy of detection of GCIs, and low rates of false acceptance and missed detection,

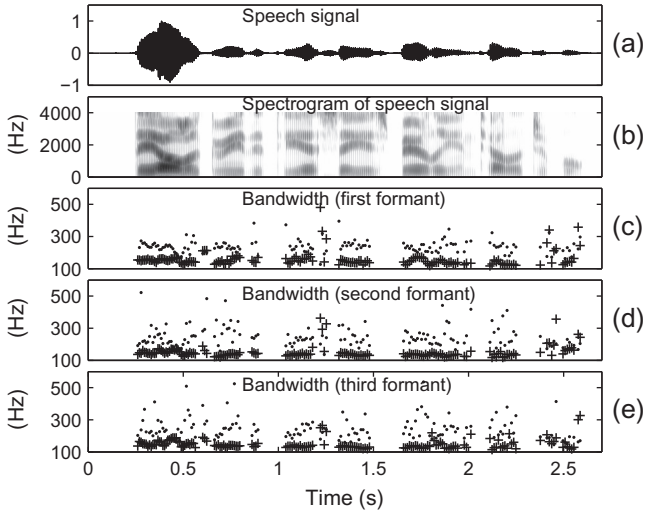


Fig. 7. (a) Speech signal and (b) its spectrogram. Bandwidths corresponding to first three formants are shown in (c), (d) and (e), respectively, at the GCIs. Bandwidths derived from regions of closed and open phase are shown by '+' and '.' symbols, respectively.

compared to other existing methods of GCI detection (Murty and Yegnanarayana, 2008). Formant frequencies are extracted from the regions of open and closed phase of glottis, using the method proposed by Xavier et al. (2006). For the speech signal in Fig. 7(a), bandwidths of the first three formants are shown in Fig. 7(c)–(e). Due to increased damping of resonances in the regions of open phase, bandwidths obtained from the open phase (shown by '.') are larger in value, compared to those obtained from the closed phase (shown by '+').

4.2.2. Method based on linear prediction analysis

In this method, signals synthesized at 10,000 Hz are processed using linear prediction (LP) analysis, which represents a signal as the output of an all-pole filter. A prediction order of 10 is used, and signal segments in closed and open regions of glottis are analyzed to extract the poles of the all-pole filter. Real poles are ignored and

Table 1
Performance of GDF-based method of formant extraction (Xavier et al., 2006). Each entry denotes the percentage of pitch periods for which the percentage deviation of the extracted formants is lesser than a certain value (such as 2%, 5% and 10%). The results are given for signals synthesized using three different values of the pitch period T_0 (5 ms, 8 ms and 10 ms). For each pitch period, formant extraction is performed in the closed phase and in the open phase.

T_0 (ms)		Closed phase			Open phase		
		$\delta_F < 2\%$	$\delta_F < 5\%$	$\delta_F < 10\%$	$\delta_F < 2\%$	$\delta_F < 5\%$	$\delta_F < 10\%$
5	F_1	81.0	91.1	94.9	37.2	80.4	94.1
	F_2	86.4	92.3	96.0	47.8	78.4	94.1
	F_3	90.5	93.1	94.1	25.9	60.7	90.8
8	F_1	94.1	98.0	98.4	58.8	93.5	98.6
	F_2	95.8	97.4	98.0	62.0	91.0	98.4
	F_3	97.2	97.6	98.3	39.4	80.9	98.2
10	F_1	98.6	99.4	99.6	76.3	98.1	99.6
	F_2	98.6	98.9	99.1	73.9	95.9	99.6
	F_3	98.9	99.1	99.5	54.0	93.9	99.4

complex-conjugate pole pairs are considered. Formant frequencies are obtained from the angle/phase of the complex-conjugate pole pairs, while bandwidths are obtained from the radial distance of the poles from the origin. Due to minimization of error in linear prediction analysis, estimates of formant frequencies and bandwidths are obtained simultaneously.

Table 2 shows the accuracy of formant extraction using linear prediction analysis. In the table, each entry denotes the percentage of pitch periods for which the percentage deviation δ_F of the extracted formant is lesser than a certain value (such as 2%, 5% and 10%). It is observed from Table 2 that the difference between the accuracy of formant extraction in closed and open phases is higher than that due to the GDF based method (Table 1). The spectral matching formulation used in LP analysis is more suitable for the closed phase regions than for the open phase regions, since the spectral peaks are smeared in the open phase due to increased damping of resonances. Again, the accuracy of formant extraction improves as the pitch period increases from 5 ms to 10 ms, since the estimates of autocorrelation sequence improve due to increase in the length of the available signal segment. From Tables 1 and 2, a comparison of the entries in the column corresponding to $\delta_F < 2\%$ for closed phase indicates that the GDF based method is more accurate for formant extraction than the method based on LP analysis. The accuracy of LP analysis for bandwidth extraction is discussed in Section 4.3.

4.2.3. Method based on filter bank analysis

We describe the method proposed by Potamianos and Maragos (1995), for extraction of bandwidths associated with resonances. A resonance $r(t)$ is isolated by filtering the input signal, and is then demodulated to obtain the constituents $a(t)$ and $f(t)$, which represent the amplitude-modulated (AM) and frequency-modulated (FM) components, respectively. A multiband demodulation analysis is used for this purpose (Tsiakoulis et al., 2013). For each filter with an impulse response $h(t)$ and center frequency f_c, an array of $2K+1$ filters is created by varying the center frequency in the vicinity of f_c as follows:

$$f_{c,k} = f_c + k\Delta f, \quad k = -K \ldots, -1, 0, 1, \ldots K, \tag{20}$$

where Δf is the distance (in frequency) between adjacent filters. A time–frequency distribution of amplitude envelopes ($a(t,k)$) and instantaneous frequency signals ($f(t,k)$) are obtained for each resonance. Each filter in the filter bank is Gabor filter. Ten filters spaced within $\pm 20\%$ of the center frequency are used for estimating the formant frequencies and their bandwidths. The method starts with a center frequency and then estimates the resonant frequency and its bandwidth. For bandwidth extraction from speech signals, the center frequencies (f_c) are chosen as the true/reference formant frequencies from VTR database. The short-time resonant frequency and bandwidth are estimated as follows:

$$F_A = \frac{\sum_k \left(\int_{t0}^{t0+T} f(t,k) a^2(t,k) dt \right)}{\sum_k \left(\int_{t0}^{t0+T} a^2(t,k) dt \right)} \tag{21}$$

$$B_A = \frac{\sum_k \left(\int_{t0}^{t0+T} [(\dot{a}(t)/2\pi)^2 + (f(t,k) - F_A)^2 a^2(t,k)] dt \right)}{\sum_k \left(\int_{t0}^{t0+T} a^2(t,k) dt \right)} \tag{22}$$

where t_0 denotes the instant of glottal closure or opening (depending on the position of the analysis window), and T denotes the duration of analysis window (half of the pitch period in this case). Synthesized signals sampled at 16,000 Hz are used for extraction of bandwidths. The accuracy of filter bank analysis for bandwidth extraction is discussed in Section 4.3.

4.3. Performance of different methods of bandwidth extraction

In this section, we discuss the performance of the three methods of bandwidth extraction, which were described in Section 4.2. We refer to the three methods as GDF, LPA and FBA. Let B denote the true value of bandwidth

Table 2
Performance of LP analysis – based method of formant extraction. Each entry denotes the percentage of pitch periods for which the percentage deviation of the extracted formants is lesser than a certain value (such as 2%, 5% and 10%). The results are given for signals synthesized using three different values of the pitch period T_0 (5 ms, 8 ms and 10 ms). For each pitch period, formant extraction is performed in the closed phase and in the open phase.

T_0 (ms)		Closed phase			Open phase		
		$\delta_F < 2\%$	$\delta_F < 5\%$	$\delta_F < 10\%$	$\delta_F < 2\%$	$\delta_F < 5\%$	$\delta_F < 10\%$
5	F_1	27.1	59.7	94.7	19.6	46.9	81.3
	F_2	68.4	93.8	99.1	25.5	54.7	83.5
	F_3	79.1	94.1	99.0	15.1	39.4	76.2
8	F_1	27.9	61.4	95.6	20.0	48.7	84.9
	F_2	71.1	95.6	99.7	20.8	49.2	82.6
	F_3	81.0	94.8	99.4	15.7	39.1	76.6
10	F_1	28.7	62.4	96.1	21.3	51.1	86.2
	F_2	71.8	96.5	99.4	21.5	47.1	80.9
	F_3	83.3	94.5	99.7	15.6	39.3	77.7

(i.e., 3 dB bandwidth in Hz) associated with a formant frequency, and let $\widehat{B}$ denote the value of bandwidth extracted from the synthetic signal. The percentage deviation δ_B is given by $\delta_B = \frac{|B-\widehat{B}|}{B} \times 100$. The accuracy of extraction of bandwidths is measured in terms of the percentage of signal frames (pitch periods) for which the deviation δ_B is lesser than a certain value (such as 5%, 10% and 20%).

Table 3 shows the performance of bandwidth extraction using the GDF method, when true/reference formant frequencies are used to extract bandwidths. Table 4 shows the performance of bandwidth extraction using the GDF method, when formant frequencies are estimated from the signals. In Tables 3 and 4, the accuracy of bandwidth extraction is better in the closed phase than in the open phase, for different pitch periods. In both cases, the accuracy of bandwidth improves as the pitch period increases. Since the formant frequencies extracted from the signal are less accurate for short windows of analysis, the accuracy of bandwidth extraction in Table 4 is somewhat poorer at 5 ms (columns $\delta_B < 5\%$ and $\delta_B < 10\%$), compared to that in Table 3. The difference between Tables 3 and 4 is more significant in open phase regions than in closed phase regions, due to poorer estimates of formant frequencies in the open phase regions. The accuracy of extraction of bandwidth is highest for third formant frequency, followed by the second and the first formant frequency. This observation holds for both Tables 3 and 4, for all the pitch periods, and for both closed and open phases. This is due to the greater bandwidth associated with the higher formant frequencies. For instance, an error of 30 Hz amounts to a deviation of 10% for a bandwidth of 300 Hz, but the same error of 30 Hz amounts to a deviation of 20% for a bandwidth of 150 Hz.

Table 5 shows the accuracy of bandwidth extraction using the LPA method. There is an improvement in the accuracy with the increase in pitch period, primarily due to improved estimates of autocorrelation coefficients from the signal. In the case of LP analysis, bandwidth is a consequence of an error minimization/ spectral matching process, which is mainly intended to match the spectral peaks of the signal with those due to the all-pole model. From Tables 4 and 5, it is observed that the GDF method has a better accuracy of bandwidth extraction compared to the LPA method, for all the three pitch periods, and in closed and open phases. The accuracy of LPA method is affected by the choice of order of LP analysis. A 10^{th} order LP analysis was used to account for the three pole-pairs used for synthesis, and to account for spectral slope. Such a choice of the order of LP analysis is suitable for most pitch periods, but the extracted formants (and hence the bandwidths) could be erroneous in some pitch periods. The extent of degradation in the accuracy from closed phase to open phase is similar in GDF and LPA methods.

Table 6 shows the accuracy of bandwidth extraction using the FBA method. This method averages the instantaneous amplitude and frequency signals across filterbank arrays. For a short window of analysis, spectral resolution using filterbank arrays is poor. Accuracy of the method improves significantly, when the pitch period increases from 5 ms to 10 ms. This method assumes an AM–FM model of speech signal. Hence, estimation of formant bandwidths is dependent on the amplitude envelopes and the instantaneous frequency signals. The method also requires some initial estimates of center frequencies for estimation of formant frequencies. In this study, we have provided the formant frequencies themselves as the initial estimates of center frequencies, so that formant extraction (and hence bandwidth extraction) is accurate. In practice, the accuracy of formant extraction has an effect on that of bandwidth extraction in the FBA method.

From Tables 5 and 6, it is observed that the FBA method has a better accuracy of bandwidth extraction compared that of the LPA method. Although both LPA and FBA are model-based methods, the averaging of instantaneous amplitude and frequency signals across filterbank arrays in the FBA method ensures better accuracy of bandwidth extraction compared to the LPA method.

Table 3
Performance of GDF method of bandwidth extraction, using true/reference values of formant frequencies. Each entry denotes the percentage of pitch periods for which the deviation of the extracted bandwidths is lesser than a certain value (such as 5%, 10% and 20%). The results are given for signals synthesized using different values of the pitch period T_0 (5 ms, 8 ms and 10 ms). For each pitch period, bandwidth extraction is performed in the closed phase and in the open phase.

T_0 (ms)		Closed phase			Open phase		
		$\delta_B < 5\%$	$\delta_B < 10\%$	$\delta_B < 20\%$	$\delta_B < 5\%$	$\delta_B < 10\%$	$\delta_B < 20\%$
5	B_1	32.0	58.3	88.1	30.4	54.7	85.5
	B_2	38.4	62.1	89.9	33.8	57.6	88.0
	B_3	53.0	81.0	96.0	52.0	80.7	97.1
8	B_1	24.5	49.5	89.6	26.4	49.1	87.4
	B_2	40.1	65.4	93.0	33.0	56.7	90.4
	B_3	53.0	81.3	97.7	43.4	72.0	95.2
10	B_1	26.8	52.1	91.8	35.4	59.2	89.2
	B_2	40.2	65.5	93.2	31.9	57.9	92.5
	B_3	51.7	79.9	97.6	43.3	71.1	95.2

Table 4
Performance of GDF method of bandwidth extraction, using estimated values of formant frequencies. Each entry denotes the percentage of pitch periods for which the deviation of the extracted bandwidths is lesser than a certain value (such as 5%, 10% and 20%). The results are given for signals synthesized using three different values of the pitch period T_0 (5 ms, 8 ms and 10 ms). For each pitch period, bandwidth extraction is performed in the closed phase and in the open phase.

T_0 (ms)		Closed phase			Open phase		
		$\delta_B < 5\%$	$\delta_B < 10\%$	$\delta_B < 20\%$	$\delta_B < 5\%$	$\delta_B < 10\%$	$\delta_B < 20\%$
5	B_1	26.5	53.7	86.8	13.7	37.3	83.9
	B_2	28.6	52.9	86.2	22.3	47.8	84.2
	B_3	39.7	72.7	94.5	32.7	64.7	91.4
8	B_1	24.0	49.1	89.5	20.8	43.3	85.4
	B_2	27.5	55.9	90.5	20.2	45.5	87.2
	B_3	35.1	68.5	95.4	37.5	68.2	94.4
10	B_1	26.6	51.9	91.4	32.5	57.1	88.6
	B_2	27.9	56.3	91.2	24.0	50.8	90.6
	B_3	34.7	68.1	95.7	36.5	66.8	94.3

Table 5
Performance of LPA method of bandwidth extraction. Each entry denotes the percentage of pitch periods for which the deviation of the extracted bandwidths is lesser than a certain value (such as 5%, 10% and 20%). The results are given for signals synthesized using three different values of the pitch period T_0 (5 ms, 8 ms and 10 ms). For each pitch period, bandwidth extraction is performed in the closed phase and in the open phase.

T_0 (ms)		Closed phase			Open phase		
		$\delta_B < 5\%$	$\delta_B < 10\%$	$\delta_B < 20\%$	$\delta_B < 5\%$	$\delta_B < 10\%$	$\delta_B < 20\%$
5	B_1	19.1	38.5	78.2	18.9	38.5	76.7
	B_2	20.5	39.5	80.3	18.2	37.3	76.0
	B_3	19.7	40.7	81.0	19.7	39.1	77.7
8	B_1	21.8	42.2	83.4	19.8	39.5	79.3
	B_2	21.3	46.9	86.8	17.1	36.7	78.5
	B_3	21.6	43.3	84.8	20.4	40.1	80.0
10	B_1	23.2	44.6	86.9	20.3	40.1	82.7
	B_2	24.3	47.6	89.6	22.6	44.1	84.7
	B_3	22.7	45.0	87.9	20.1	41.0	83.3

Table 6
Performance of FBA method of bandwidth extraction. Each entry denotes the percentage of pitch periods for which the deviation of the extracted bandwidths is lesser than a certain value (such as 5%, 10% and 20%). The results are given for signals synthesized using three different values of the pitch period T_0 (5 ms, 8 ms and 10 ms). For each pitch period, bandwidth extraction is performed in the closed phase and in the open phase.

T_0 (ms)		Closed phase			Open phase		
		$\delta_B < 5\%$	$\delta_B < 10\%$	$\delta_B < 20\%$	$\delta_B < 5\%$	$\delta_B < 10\%$	$\delta_B < 20\%$
5	B_1	19.8	38.0	77.0	15.0	32.5	74.0
	B_2	21.3	42.4	80.7	14.0	29.7	76.2
	B_3	33.6	59.2	87.1	12.9	28.2	74.8
8	B_1	19.9	40.7	82.1	18.8	36.8	79.4
	B_2	29.4	56.1	91.8	20.4	40.9	83.7
	B_3	33.6	59.2	89.8	13.0	28.9	80.4
10	B_1	22.7	45.1	87.8	15.0	31.7	81.9
	B_2	36.2	63.4	94.3	23.7	46.5	89.2
	B_3	35.2	61.6	91.6	13.7	30.2	83.5

The LPA method is essentially an error minimization process which does not actively search for spectral peaks of the signal. Both LPA and FBA methods are sensitive to the choice of model parameters. The GDF method is partly model-based, in the sense that it relies on the equivalence of the theoretical GDF of an all-pole filter and the GDF computed from the speech signal. A comparison of Tables 4 and 6 indicates that the accuracy of the GDF method is better than that of the FBA method. In the case of GDF and FBA methods, formant tracking can be employed to

exploit the temporal continuity of formant frequencies, to improve the accuracy of formant extraction. This can help in improving the accuracy of bandwidth extraction.

5. Conclusion

A new method for extraction of bandwidths of formant frequencies is proposed in this paper, based on the properties of group delay function (GDF). The method exploits the presence of prominent peaks in the GDF at resonant frequencies, and the relatively less influence of one resonant peak on the other peaks in GDF. An all-pole model of vocal tract system was assumed to illustrate the basis for the method. Accuracy of the method for extraction of bandwidths was evaluated for synthesized signals. An important feature of the method is that it can extract bandwidths from short segments of signals. Accuracy of the method is demonstrated by comparing with two existing methods of bandwidth extraction. In practice, the effectiveness of the GDF-based method depends on the accuracy of extraction of formant frequencies, and also on the pitch period available for analysis.

References

Cohen, L., Assaleh, K., Fineberg, A., 1992. Instantaneous bandwidth and formant bandwidth. In: Proceedings of the IEEE Sixth SP Workshop on Statistical Signal and Array Processing, Victoria, BC, Canada, pp. 13–17.

Deng, L., Cui, X., Pruvenok, R., Huang, J., Momen, S., Chen, Y., Alwan, A., 2006. A database of vocal tract resonance trajectories for research in speech processing. In: Proceedings of the International Conference on Acoustics, Speech Signal Processing. Toulouse, France, pp. 369–372.

Fant, G., 1960. Acoustic Theory of Speech Production. Mouton, The Hague, Netherlands.

Gauffin, J., Sundberg, J., 1989. Spectral correlates of glottal voice source waveform characteristics. J. Speech. Hear. Res. 32 (3), 556–565.

Makhoul, J., 1975. Linear prediction: a tutorial review. In: Proceedings of the IEEE, vol. 63(4), pp. 561–580.

Murty, K.S.R., Yegnanarayana, B., 2008. Epoch extraction from speech signals. IEEE Trans. Audio, Speech Lang. Process. 16 (8), 1602–1613.

Oppenheim, A.V., Schafer, R.W., 1975a. Digital Signal Processing. Prentice Hall, Englewood Cliffs, New Jersey (Ch. 10).

Oppenheim, A.V., Schafer, R.W., 1975b. Digital Signal Processing. Prentice Hall, Englewood Cliffs, New Jersey.

Oppenheim, A.V., Schafer, R.W., Buck, J.R., 1999. Discrete-Time Signal Processing. Prentice Hall, Upper Saddle River, New Jersey (Ch. 5).

Potamianos, A., Maragos, P., 1995. Speech formant frequency and bandwidth tracking using multiband energy demodulation. In: Proceedings of the International Conference on Acoustics, Speech Signal Process, vol. 1. Detroit, MI, USA, pp. 784–787.

Reddy, N., Swamy, M., 1984. High resolution formant extraction from linear-prediction phase spectra. IEEE Trans. Acoust. Speech Signal Process. 32 (6), 1136–1144.

Tsiakoulis, P., Potamianos, A., Dimitriadis, D., 2013. Instantaneous frequency and bandwidth estimation using filterbank arrays. In: Proceedings of the International Conference on Acoustics, Speech Signal Process. Vancouver, BC, pp. 8032–8036.

Xavier, M.A.J., Guruprasad, S., Yegnanarayana, B., 2006. Extracting formants from short segments of speech using group delay functions. In: Proceedings of the INTERSPEECH. Pittsburgh, PA, USA, pp. 1009–1012.

Yasojima, O., Takahashi, Y., Tohyama, M., 2006. Resonant bandwidth estimation of vowels using clustered-line spectrum modeling for pressure speech waveforms. In: Proceedings of the IEEE International Symposium on Signal Processing and Information Technology. Vancouver, Canada, pp. 589–593.

Yegnanarayana, B., 1978. Formant extraction from linear prediction phase spectra. J. Acoust. Soc. Am. 63 (5), 1638–1640.

Zheng, Y., Hasegawa-Johnson, M., 2003. Particle filtering approach to Bayesian formant tracking. In: Proceedings of the IEEE Workshop on Statistical Signal Processing. St. Louis, MO, USA, pp. 601–604.

Query-by-Example Spoken Term Detection using Frequency Domain Linear Prediction and Non-segmental Dynamic Time Warping

Gautam Mantena, Sivanand Achanta and Kishore Prahallad

***Abstract*—The task of query-by-example spoken term detection (QbE-STD) is to find a spoken query within spoken audio data. Current state-of-the-art techniques assume zero prior knowledge about the language of the audio data, and thus explore dynamic time warping (DTW) based techniques for the QbE-STD task. In this paper, we use a variant of DTW based algorithm referred to as non-segmental DTW (NS-DTW), with a computational upper bound of $O(mn)$ and analyze the performance of QbE-STD with Gaussian posteriorgrams obtained from spectral and temporal features of the speech signal. The results show that frequency domain linear prediction cepstral coefficients, which capture the temporal dynamics of the speech signal, can be used as an alternative to traditional spectral features such as linear prediction cepstral coefficients, perceptual linear prediction cepstral coefficients and Mel-frequency cepstral coefficients.**

We also introduce another variant of NS-DTW called fast NS-DTW (FNS-DTW) which uses reduced feature vectors for search. With a reduction factor of $\alpha \in \mathbb{N}$, we show that the computational upper bound for FNS-DTW is $O\left(\frac{mn}{\alpha^2}\right)$ which is faster than NS-DTW.

***Index Terms*—Query-by-example spoken term detection, dynamic time warping, fast search, frequency domain linear prediction.**

EDICS Category: SLP-SMIR

I. Introduction

The task of query-by-example spoken term detection (QbE-STD) is to find a spoken query within spoken audio. A key aspect of QbE-STD is to enable searching in multi-lingual and multi-speaker audio data. A traditional QbE-STD approach is to convert spoken audio into a sequence of symbols and then perform text based search. In [1]–[3], the audio is first converted to a sequence of symbols using automatic speech recognition (ASR) and then lattice based search techniques are incorporated.

ASR based techniques assume the availability of labelled data for training the acoustic and language models. Such approaches are not scalable for languages where there is no availability or the resources to build an ASR. To overcome this limitation, zero prior knowledge is assumed about the language of the spoken audio, and thus dynamic time warping (DTW) based techniques are exploited for QbE-STD [4]–[9]. One of the popular DTW based techniques is the segmental DTW (S-DTW) [4], which uses a windowed (or segmental) type of approach to search a spoken query within spoken audio. In this paper, we use a variant of DTW referred to as non-segmental DTW (NS-DTW) which has been applied for segmentation of large speech files [10], [11] and also for QbE-STD tasks [6], [8], [9]. In [12], the NS-DTW is referred to as subsequence DTW.

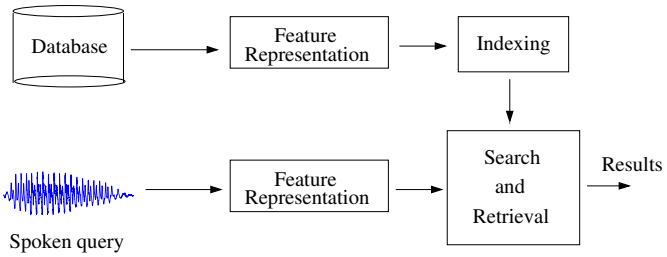


Fig. 1. A general architecture for a QbE-STD system.

Fig. 1 shows a general architecture of a QbE-STD system. Speech features are extracted from the audio database and are indexed for quick retrieval during the search process. In [4], [8], [13], Gaussian posteriorgrams are shown to be a good feature representation to suppress speaker characteristics and to perform search across multi-lingual data.

In general, Gaussian posteriorgrams used for QbE-STD are computed from short-time spectral features such as Mel-frequency cepstral coefficients. In [14], [15], it is shown that frequency domain linear prediction cepstral coefficients (FDLP) perform better than the short-time spectral features for speech recognition in noisy environments. In FDLP, the temporal dynamics of the speech signal are captured by applying an all-pole model in the spectral domain. Athineos *et. al.* [16] provides a detailed mathematical analysis of extracting the temporal envelope of the signal using autoregressive modelling. In this paper, we show that Gaussian posteriorgrams computed from FDLP, which capture the temporal dynamics of the speech signal, can be used as an alternative to traditional spectral parameters such as linear prediction cepstral coefficients (LPCC), perceptual linear prediction cepstral coefficients (PLP) and Mel-frequency cepstral coefficients (MFCC).

In [7], [17], indexing based approaches such as locality sensitive hashing and hierarchical clustering are used to build sparse similarity matrices for searching the spoken query. Use of indexing techniques is not in the scope of this work. Thus a spoken utterance is represented by a sequence of Gaussian posteriorgrams and a full similarity matrix is used for searching a spoken query.

The brief set of contributions of our work is as follows:

- We provide a comparison of time complexity of NS-

DTW and S-DTW [4]. We experiment with different local constraints in NS-DTW based method, and report results on the common MediaEval 2012 dataset [18].
- In this work, we introduce a faster method of searching a spoken query. This method exploits the redundancy in speech signal, and averages the successive Gaussian posteriorgrams to reduce the length of the spoken audio and the spoken query. However with such an approach there is a trade-off between search performance and accuracy and these results are reported. We show that the search time of the proposed fast NS-DTW is lower than that of the randomized acoustic indexing method described in [19].
- We provide experimental results to show that the Gaussian posteriorgrams obtained from FDLP can be used for QbE-STD as an alternative to other short-time spectral features such as MFCC.

II. Database

The experiments conducted in this work use MediaEval 2012 data which is a subset of Lwazi database [18]. The data consists of audio recorded via telephone in 4 of 11 South African languages. We have considered two data sets, development (dev) and evaluation (eval) which contain spoken audio (reference) and spoken query data. The statistics of the audio data is shown in Table I.

TABLE I
STATISTICS OF MEDIAEVAL 2012 DATA.

Data	Utts	Total(min)	Average(sec)
dev reference	1580	221.9	8.42
dev query	100	2.4	1.44
eval reference	1660	232.5	8.40
eval query	100	2.5	1.50

III. Feature Representation for Speech

Feature representation of the speech signal was obtained by a two step process. In the first step, parameters were extracted from the speech signal. In the second step, Gaussian posteriorgrams were computed from these parameters. The different parameters extracted from the speech signal were as follows: (a) Linear prediction cepstral coefficients (LPCC), (b) Mel-frequency cepstral coefficients (MFCC), (c) Perceptual linear prediction cepstral coefficients (PLP) [20] and (d) Frequency domain linear prediction cepstral coefficients (FDLP).

$$X[k] = a[k] \sum_{n=0}^{N-1} x[n] \; cos\left(\frac{(2n+1)\pi k}{2N}\right) \quad k = 0, 1, ..., N-1$$

where:

$$a[k] = \begin{cases} \frac{1}{\sqrt{N}} & k = 0 \\ \sqrt{\frac{2}{N}} & k = 1, 2, ..., N-1 \end{cases} \quad (1)$$

In linear prediction (LP) analysis of speech an all-pole model was used to approximate the vocal tract spectral envelope [21]. MFCC, PLP and FDLP use a series of band pass filters to capture speech specific characteristics. To compute MFCC, signal was passed through a bank of filters to compute the energy of the signal in each of the bands. This energy from each band is referred to as Mel-spectrum. Cepstral coefficients were then computed by performing DCT on these sub-band energies. In PLP analysis of speech, the power spectrum was modified before applying the LP all-pole model. The modified spectrum was obtained as follows [20]: (a) speech signal is first passed through the filter banks, (b) pre-emphasis by an equal loudness curve on the filtered signal and (c) cubic compression of the spectrum.

In LPCC, MFCC and PLP the short-time spectral properties of the speech signal are captured. In order to capture the temporal dynamics of the speech signal, frequency domain linear prediction (FDLP) was developed [14]–[16]. FDLP technique relies on all-pole modeling in the spectral domain to characterize the temporal dynamics of the frequency components. In [15], the performance of FDLP parameters for phoneme recognition was evaluated in noise conditions such as additive noise, convolutive noise and telephone channel. It was shown that, in such noise conditions, FDLP was performing better as compared to other parameters such as PLP. This motivated us to explore FDLP based features for QbE-STD.

Following the work in [22], FDLP parameters were computed as follows – (a) DCT was computed over the entire signal using Eq. (1), (b) Filter bank analysis was performed on the DCT output. (c) An all-pole model was applied on the spectral components in each sub-band, (d) For each sub-band, time domain envelope was computed by taking the frequency response of the all-pole model, (e) Short-time analysis was performed on the envelopes from each of the sub-bands to compute the FDLP spectrum, and (f) DCT was then applied on the FDLP spectrum to obtain cepstral coefficients.

A. Representation using Gaussian Posteriorgrams

Gaussian posteriorgrams were computed from the 39 dimensional LPCC, MFCC, PLP and FDLP parameters. A 25 ms window length with 10 ms shift was considered to extract 13 dimensional parameters along with delta and acceleration coefficients for all the parameters. An all-pole model of order 12 was used for LPCC, PLP and an order of 160 poles/sec for the FDLP parameters. A set of 26 filter banks were used for computing MFCC, PLP and 37 filter banks for the FDLP parameters.

Gaussian posteriorgrams were computed from these parameters as described in [8]:

1) K-means was used to initialize the means of the Gaussian mixture models (GMM). The initialization started by computing the mean μ and standard deviation σ from the entire data. Then a split operation was performed and the new centers were given by $\mu \pm 0.2\sigma$. The process of clustering and splitting continued till the required number of means were reached.
2) GMMs were trained with its centers initialized by K-means.

3) As a final step, feature vectors were pooled for each Gaussian having a maximum likelihood and the means and covariances were recomputed.

IV. Search using Non-segmental DTW

Dynamic time warping (DTW) algorithm performs a non-linear alignment of two time series. During this process, the warping constraints such as (1) start and end point, (2) monotonicity, (3) local, (4) global and (5) slope weighting are considered [23].

In segmental DTW (S-DTW), we use global constraints to restrict the alignment within a certain segment of the spoken audio. Segmenting the spoken audio using global constraints and then performing DTW is computationally expensive. As an alternative, we use non-segmental DTW (NS-DTW), where we approximate the start and end point constraints.

Let $\mathcal{Q}$ be a spoken query (or query) containing n feature vectors. Let $\mathcal{R}$ be the spoken audio (or reference) containing m feature vectors. The sequence of feature vectors are denoted as follows:

$$\mathcal{Q} = \{\mathbf{q_1}, \mathbf{q_2}, \ldots, \mathbf{q_i}, \ldots, \mathbf{q_n}\},$$
$$\mathcal{R} = \{\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_j}, \ldots, \mathbf{u_m}\}.$$

Each of these feature vectors represent a Gaussian posteriorgram as computed in Section III-A. The distance measure between a query vector $\mathbf{q_i}$ and a reference vector $\mathbf{u_j}$ is given by Eq. (2)

$$d(i,j) = -log\left(\frac{\mathbf{q_i}}{||\mathbf{q_i}||} \cdot \frac{\mathbf{u_j}}{||\mathbf{u_j}||}\right) \tag{2}$$

We define the term *search hit* as the region in the reference $\mathcal{R}$ that is likely to contain the query $\mathcal{Q}$. In NS-DTW, we use only the local constraints as shown in Fig. 2 to obtain the *search hits*. The choice of these local constraints is motivated by their use in isolated word recognition [24] and in large vocabulary speech recognition [25], [26]. These local constraints are often referred as Bakis topology [25]. In Section V-D, we compare the performance of different sets of local constraints for QbE-STD tasks.

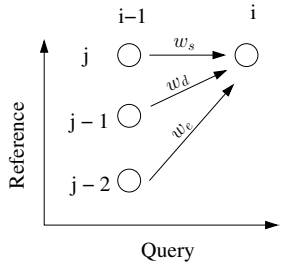


Fig. 2. A pictorial representation of the local constraints along with the weights w_s, w_d and w_e associated with each of the arcs.

We compute a similarity matrix S of size $m \times n$, where m, n are the number of feature vectors of the reference and the query. The query can start from any point in the reference. Initially, $S(1,j) = d(1,j)$, where $d(1,j)$ is the distance measure given by Eq. (2). The entries in the rest of the similarity matrix is given by Eq. (3) [8].

$$S(i,j) = min\left\{\begin{array}{c}\frac{d(i,j)+S(i-1,j-2)}{T(i-1,j-2)+w_e}\\ \frac{d(i,j)+S(i-1,j-1)}{T(i-1,j-1)+w_d}\\ \frac{d(i,j)+S(i-1,j)}{T(i-1,j)+w_s}\end{array}\right\}, \tag{3}$$

where T is called the transition matrix. $T(i,j)$ represents the number of transitions required to reach i,j from a start point, and normalizes the accumulated score with the length of the aligned path. The update equation for the transition matrix T is given by Eq. (4).

$$T(i,j) \quad = \quad \left\{\begin{array}{lll} T(i-1,\hat{j}) + w_e & \text{if} & \hat{j} = j-2 \\ T(i-1,\hat{j}) + w_d & \text{if} & \hat{j} = j-1 \\ T(i-1,\hat{j}) + w_s & \text{if} & \hat{j} = j \end{array}\right. \tag{4}$$

where

$$\hat{j} \quad = \quad \underbrace{argmin}_{\hat{j}\in\{j,\ j-1,\ j-2\}} \left\{\begin{array}{c}\frac{d(i,j)+S(i-1,j-2)}{T(i-1,j-2)+w_e}\\ \frac{d(i,j)+S(i-1,j-1)}{T(i-1,j-1)+w_d}\\ \frac{d(i,j)+S(i-1,j)}{T(i-1,j)+w_s}\end{array}\right\}.$$

In Eq. (4), w_e, w_d, w_s are the weights associated for each transition. In Section V-A, we show the effect of weights on the search performance of NS-DTW and thereby select the optimum values for the weights.

A. Selection of Start and End Time Stamps

In order to detect the start and end time stamps of the *search hit*, we obtain the reference index that contains the best alignment score, i. e., the end point of the *search hit* as given by $j = \min_j\{S(n,j)\}$ for $j = 1, 2, ..., m$. Once the end point j is obtained, the corresponding start point could be obtained by using a token passing algorithm as shown in Eq. (5) and Eq. (6).

$$P(i,1) \quad = \quad i \quad \text{for} \quad i = 1,2,3,...,m \tag{5}$$
$$P(i,j) \quad = \quad P(i-1,\hat{j}), \tag{6}$$

where P is a matrix to record the path transitions. The values in the matrix $P(i,j)$ are updated when the similarity matrix is being computed and thus avoiding the need for path traceback to obtain the start time stamp of the search hit.

Fig. 3(a) shows an example similarity matrix plot of a query and a reference where the dark bands represent the segments that are similar between the query and the reference. To visualize the similarity matrix, each value in the matrix is scaled using an exponential function and then each column is normalized by the maximum value of the column. Please note that a full similarity matrix is computed and the white regions (as shown in Fig. 3) does not imply that we do not compute the values of the matrix in those regions.

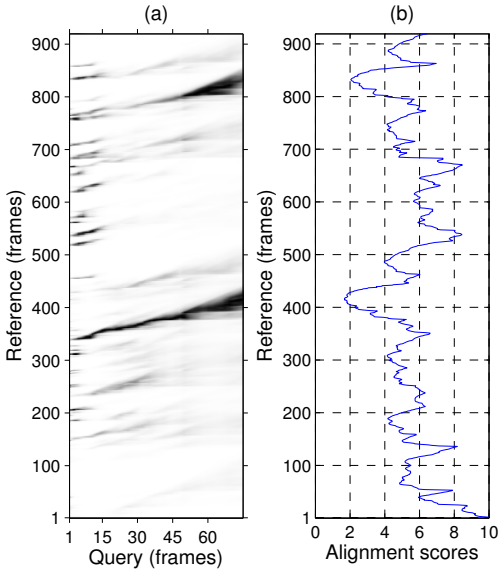


Fig. 3. (a) An example similarity matrix plot obtained using NS-DTW when a query is present in the reference and (b) A plot of the alignment scores obtained from the last column of the similarity matrix. Please note that, to visualize the similarity matrix, the values in the matrix are scaled using an exponential function and then each column is normalized with the maximum value of the column.

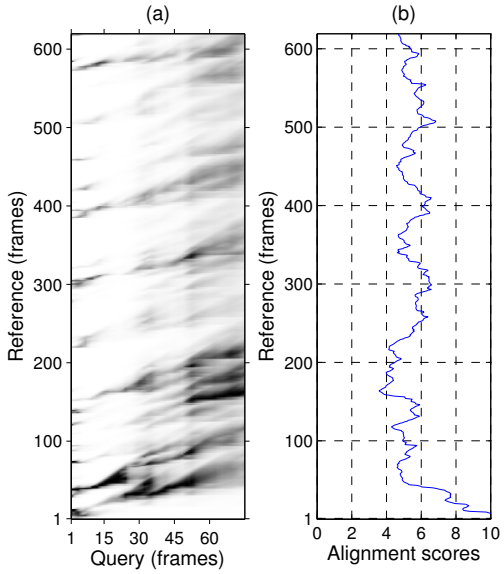


Fig. 4. (a) An example similarity matrix plot obtained using NS-DTW when a query is not present in the reference and (b) A plot of the alignment scores obtained from the last column of the similarity matrix. Please note that, to visualize the similarity matrix, the values in the matrix are scaled using an exponential function and then each column is normalized with the maximum value of the column.

The dark bands that have reached the end of the query are the required *search hits*. They can be obtained from the alignment scores from the last column of the similarity matrix S. Fig. 3(b) shows the alignment scores where the minimum values represent the end of the *search hits* and from these points the start time stamps are obtained using Eq. (5) and Eq. (6).

As shown in Fig. 3(a), the query could have more than one match in the reference and hence k-best alignment scoring indices are selected from the similarity matrix. In Section V-B, we show the effect of the choice of k-best alignment scores on the search performance of NS-DTW and thereby select the optimum k value.

Fig. 4(a) shows an example similarity matrix plot when a query is not present in the reference. The partial bands that are observed in Fig. 4(a) show a partial match between the query and the reference. From Fig. 4(b), it can be seen that the alignment scores of the search hits are higher than that of the scores of a search hit shown in Fig. 3(b).

B. Analytical Comparison with Segmental-DTW

Segmental DTW (S-DTW) [4] is a popular technique that overcomes the start and end point constraints by dividing the spoken audio into a series of segments, and then DTW is performed on each segment. S-DTW is computationally not efficient due to this segment based DTW approach that it performs to obtain the *search hits*.

Two constraints are imposed on the alignment. The first one is a parameter r which dictates the length of the segment to be taken from the reference. This is given by the inequality $|i-j| \leq r$ (Sakoe-Chiba band [23]), where i, j are the frame indices of the query and the reference. This constraint prevents the warping from going too far ahead or behind.

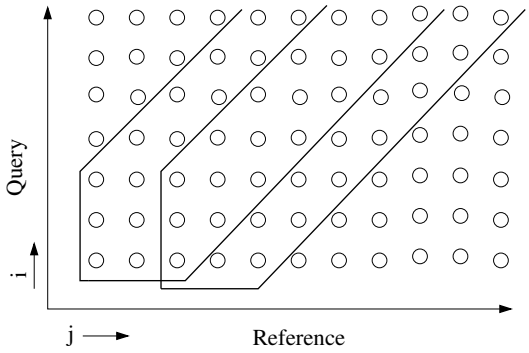


Fig. 5. An example of segmental DTW (S-DTW) with the first two segments for $r = 2$.

The second constraint is the number of such segments to be considered. Fig. 5 shows the first two segments of S-DTW for $r = 2$. Normally one would shift the segment by one frame as the query could start from any point in the reference, but due to the huge computational overload a shift of r is considered.

The total number of computations required is equal to *number of computations in each segment* $\times$ *number of segments*. Given a query $\mathcal{Q}$ of size n, the length of the segment taken from reference $\mathcal{R}$ is $n + r$ ($\because j \leq i + r$). Thus the number of computations required in

each segment is of the order $O(n^2)$. For $r = 1$, searching in a reference of size m, we need to initiate m DTW searches each of order $O(n^2)$. The overall computation would be of the order $O(mn^2)$.

In NS-DTW, we are computing a similarity matrix of size $m \times n$ and so the upper bound of NS-DTW would be $O(mn)$. This is computationally faster than the S-DTW whose upper bound is $O(mn^2)$. The upper bound on the distance computation between two vectors is $O(d)$, where d is the dimensions of the vector. This distance computation is common across S-DTW and NS-DTW and so it is omitted for calculating the computational upper bound.

In NS-DTW, in order to avoid path traceback to obtain the start and end time stamps, we use a matrix P (as given by Eq. (6)). However, one can always use path traceback for obtaining the start time stamp. In such a case, the total time complexity of searching using NS-DTW is $O(mn)+O(n)$, where $O(n)$ is time complexity of path traceback. With $m >> n$, $O(mn) + O(n) = O(mn)$. Thus, the time complexity of NS-DTW is $O(mn)$ irrespective of whether path traceback or a matrix P is used. It is to be noted that the use of a matrix P will result in a higher memory requirement for computation.

C. Variants of NS-DTW

In [6], [8], [9], variants of NS-DTW are used for QbE-STD. These variants differ in the type of local constraints, values of weights and frame-based normalization. In [6], frame-based normalization is used by dividing the values in the column by the maximum value of the column. In this work, we do not perform frame-based normalization. However, we normalize each value in the similarity matrix, $S(i, j)$, by a transition matrix value, $T(i, j)$ (as given by Eq. (3)). Further details of our implementations are described in Section V.

V. Evaluation and Results

All the evaluations are performed using 2006 NIST evaluation criteria [27] and the corresponding maximum term weighted values (MTWV) are reported. To compute the MTWV, the average miss probability (MP) and false alarm probabilities (FAP) are computed for all the queries. More details on the evaluation can be found in [28].

A. Weights of Local Constraints

As given by Eq. (3) and Eq. (5), we use weights for each of the local constraints to normalize the scores. During alignment, many deletions and insertions are an indication of the mismatch between the two sequence of feature vectors and hence more importance is given to the diagonal transition (w_d). Fig. 6 shows MTWV for various values of w_d (with $w_e = w_s = 1$). NS-DTW is evaluated using 128 dimensional Gaussian posteriorgrams computed from LPCC, PLP, MFCC and FDLP. From Fig. 6, it can be seen that (a) MFCC and FDLP based features have similar MTWV for $w_d = 3$ on the dev dataset, and (b) NS-DTW performs best for FDLP at $w_d = 2$. For all of the experiments reported in this work, $w_d = 2$ is considered based on the performance of Gaussian posteriorgrams of FDLP.

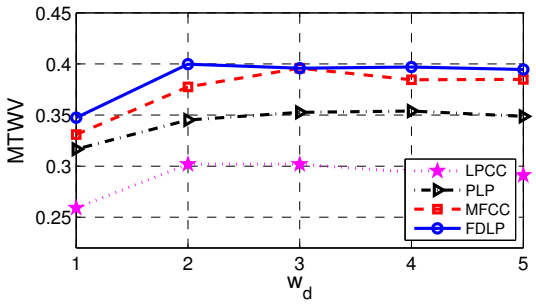


Fig. 6. Maximum term weighted value (MTWV) obtained using various values of w_d for dev dataset.

B. Selection of Number of Search Hits

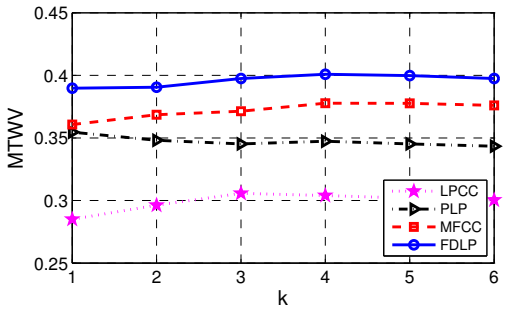


Fig. 7. MTWV obtained using various values of k using dev dataset.

In NS-DTW, after computing the similarity matrix, we select k-best alignment score indices. Using matrix P (as described in Section IV), we obtain the start time stamps of the search hits given k-best indices. After obtaining the k-best search hits, a post processing step is performed on the overlapping search hits. If there is an overlay of more than 50% between any two search hits, the search hit with the best alignment score is considered.

In a reference, there might be a possibility of multiple occurrences of the query. In such a case, $k = 1$ will result in an increase in miss probability. On the other hand a large value of k will increase in the number of false alarms. Thus, an appropriate value of k is needed. Fig. 7 shows the performance of NS-DTW for different values of k on dev dataset across different parameters. From Fig. 7, it can be seen that the MTWVs are similar for various values of k and thus the choice of $k = 5$ is chosen.

C. Number of Gaussians

Table II shows the MTWV and the search speed (in minutes) obtained using LPCC, PLP, MFCC and FDLP parameters by varying the number of Gaussians for the dev dataset. In Table II, we show the rate of improvement in the MTWV (indicated within the brackets for each of the MTWV values) by increasing the number of Gaussians. For example, the rate of improvement in MTWV for FDLP by increasing the number of Gaussians from 64 to 128 is 0.050.

TABLE II
MAXIMUM TERM WEIGHTED VALUE (MTWV) AND SEARCH SPEED[1] ON DEV DATASET BY VARYING THE NUMBER OF GAUSSIANS FOR EACH OF THE PARAMETERS. THE VALUES INDICATED IN THE BRACKETS SHOW THE RATE OF IMPROVEMENT IN THE MTWV ON INCREASING THE NUMBER OF GAUSSIANS.

No. of Gaussians	MTWV				Search Speed (mins)
	LPCC	PLP	MFCC	FDLP	
8	0.031 (-)	0.080 (-)	0.059 (-)	0.084 (-)	18.39
16	0.127 (0.096)	0.128 (0.048)	0.119 (0.06)	0.207 (0.123)	22.57
32	0.218 (0.091)	0.271 (0.143)	0.246 (0.127)	0.292 (0.085)	30.60
64	0.252 (0.034)	0.319 (0.048)	0.347 (0.101)	0.349 (0.057)	47.53
128	**0.301 (0.049)**	**0.345 (0.026)**	**0.377 (0.030)**	**0.399 (0.050)**	**80.24**
256	0.310 (0.009)	0.387 (0.042)	0.410 (0.033)	0.410 (0.011)	145.07
512	0.311 (0.001)	0.399 (0.012)	0.410 (0.000)	0.422 (0.012)	274.98
1024	0.319 (0.008)	0.404 (0.005)	0.413 (0.003)	0.432 (0.010)	534.15

In Table II, we also show the search speed, i.e. the time required to search all the queries within the dataset. The distance computation, given by Eq. (2), between a query feature (q_i) and a reference feature (u_j) is $O(d)$, where d is the dimension of the feature. This distance computation is common across S-DTW and NS-DTW and so it is omitted for calculating the computational upper bound. However, the feature dimension has an impact on the search speed of NS-DTW and is shown in Table II. The search speed of NS-DTW using a d dimensional Gaussian posteriorgram will be similar irrespective of the parameters (such as MFCC, FDLP) used to build a GMM. Thus, we have reported the search speed of NS-DTW using Gaussian posteriorgrams of FDLP by varying the number of Gaussians (as shown in Table II).

From the MTWV reported in Table II, it can be seen that (a) The performance of NS-DTW improves by increasing the number of Gaussians. However, the rate of improvement in performance for NS-DTW decreases when the number of Gaussians exceeds 128, (b) With the increase in number of Gaussians, MTWV of FDLP, MFCC and PLP seems to be converging, and (c) FDLP performs similar to that of MFCC for 256 Gaussians.

From Table II, it can also be seen that there is a trade-off between the performance of NS-DTW and the search speed by increasing the number of Gaussians. Considering the MTWV and the search speed on dev dataset we have chosen 128 Gaussians as an optimum number for NS-DTW.

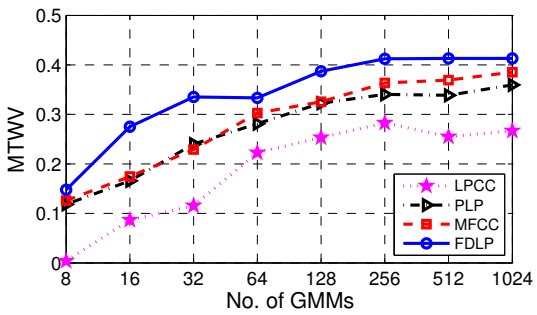


Fig. 8. Maximum term weighted values (MTWV) on eval dataset by varying the number of Gaussians for each of the parameters.

Although we have chosen 128 to be the optimum number of Gaussians, we would want to verify the effect of search performance on the eval dataset by varying the number of Gaussians. Fig. 8 shows the performance of NS-DTW using different number of GMMs trained with LPCC, MFCC, PLP and FDLP parameter streams using the eval dataset. In Fig. 8, we observe the following: (a) The curve flattens after 256 Gaussians for the features obtained from FDLP. Thus there is no further improvement in the search performance by increasing the number of Gaussians, (b) FDLP is performing better than the other acoustic parameters such as LPCC, PLP and MFCC. However, on increasing the number of Gaussians, the MTWVs of MFCC and PLP seems to be converging towards that of FDLP, and (c) Drop in the search performance for LPCC at 512 Gaussians which may be an indication of model over-fitting.

D. Effect of Different Local Constraints

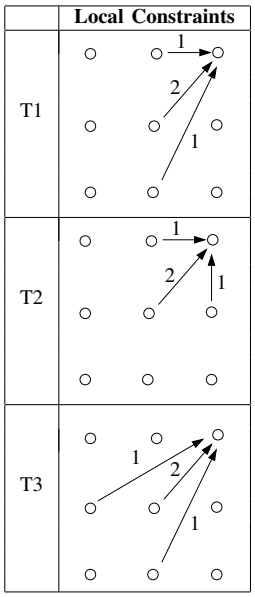
In this section, we analyse the performance of DTW-based techniques with other local constraints as shown in Table III. In [6], local constraints T2 and in [8], [9], local constraints T3 are used for QbE-STD.

Fig. 9(a) and 9(b) show the MTWV obtained using 128 dimensional Gaussian posteriorgrams of LPCC, PLP, MFCC and FDLP parameters for dev and eval datasets using T1, T2 and T3 local constraints. T1 is the local constraints used in NS-DTW (also shown in Fig. 2).

From Fig. 9(a), T2 is performing better than the other local constraints on the dev dataset. In Fig. 9(b), it can be seen that T1 is performing similar to that of T2 on eval dataset. T2 allows insertions in a query which can be interpreted as a deletion operation on the reference and this might be the reason for T1 and T2 to perform similarly on eval dataset. However, the results are not consistent, i.e., T2 performs better than T1 on dev dataset (as shown in Fig. 9). One could argue that T2 allows insertions within a query and thus more suitable for QbE-STD. As described in Section IV, we are motivated to use T1 for NS-DTW by their use in large vocabulary speech recognition and feasibility in usage of embedded training for unsupervised acoustic models with left-to-right Bakis topology [29], [30].

[1]Please note that, for a given number of Gaussians, the search speed (in NS-DTW) will be similar for each of the parameters. Thus, the search speed is reported using Gaussian posteriorgrams of FDLP.

TABLE III
SOME OF THE TYPES OF LOCAL CONSTRAINTS USED IN DTW-BASED QBE-STD.



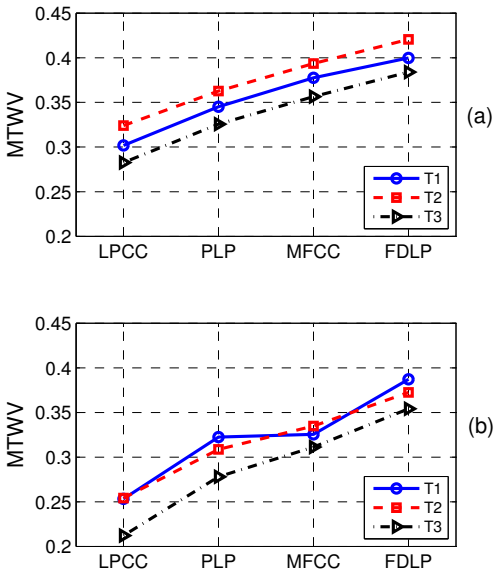


Fig. 9. MTWV obtained using 128 dimensional Gaussian posteriorgrams of various parameters using T1, T2 and T3 local constraints for (a) dev, and (b) eval datasets.

E. Use of FDLP for QbE-STD

Speech parameters such as LPCC, PLP and MFCC are obtained by windowing the speech signal and followed by estimating the spectrum from each window. However, speech signal has information spread across longer temporal context and this information can be captured by using FDLP parameters. In Table II, it can be seen that FDLP performs similar to that of MFCC using 256 Gaussians. Thus, we show that FDLP parameters, which capture the temporal characteristics of a speech signal, can be used as an alternative to other spectral parameters such as MFCC. In Fig. 8, it can be seen that FDLP performs better than MFCC for 128 and 256 GMMs and thus a motivation to use FDLP parameters for QbE-STD. To summarize the search performance of the various parameters, in Table IV we show detail results in terms of MP, FAP and MTWV using 128 dimensional Gaussian posteriorgrams.

TABLE IV
MISS PROBABILITY (MP), FALSE ALARM PROBABILITY (FAP) AND MAXIMUM TERM WEIGHTED VALUE (MTWV) OBTAINED FOR NS-DTW USING GAUSSIAN POSTERIORGRAMS OBTAINED FROM LPCC, MFCC, PLP AND FDLP.

Feats.	**dev**			**eval**		
	MP	FAP (10^{-2})	MTWV	MP	FAP (10^{-2})	MTWV
LPCC	0.575	0.802	0.301	0.564	1.529	0.253
MFCC	0.492	0.848	0.377	0.572	0.860	0.325
PLP	0.504	0.982	0.345	0.505	1.441	0.322
FDLP	0.426	1.136	**0.399**	0.402	1.766	**0.387**

VI. FAST NS-DTW

The computational analysis shown in Section IV indicates that NS-DTW is faster, than S-DTW, with an upper bound of $O(mn)$. Even with this computational improvement, DTW based techniques are still slow as compared to other model based techniques [1]–[3].

Some of the standard techniques to improve the computational performance of DTW are [31]:

- *Constraints*: Use of constraints such as Sakoe-Chiba band [23] or Itakura parallelogram [24] to limit the number of computations in the similarity matrix.
- *Data Abstraction*: Use a reduced feature representation to perform DTW. To improve the computational performance of NS-DTW, we use reduced Gaussian posteriorgrams to perform the search.
- *Indexing*: Indexing based techniques retrieve the reference feature vectors used to construct a sparse similarity matrix, which makes the search efficient [7], [17]. Use of indexing techniques is not in the scope of this paper and we compute a full similarity matrix to perform the search.

In this section, we introduce a modification to NS-DTW by reducing the query and reference Gaussian posteriorgram vectors before performing search. We refer to this algorithm as fast NS-DTW (FNS-DTW). Given a reduction factor $\alpha \in \mathbb{N}$, a window of size α is considered over the posteriorgram features and a mean is computed. The window is then shifted by α and another mean vector is computed. The posteriorgram vectors are replaced with the reduced number of posteriorgram features during this process. With a reduction factor of α, the new size of the query and the reference would be $\frac{n}{\alpha}$ and $\frac{m}{\alpha}$ respectively. This would result in a computational upper bound of $O\left(\frac{mn}{\alpha^2}\right)$ for FNS-DTW. This technique is independent of the local constraints used and we use T1 local constraints for FNS-DTW.

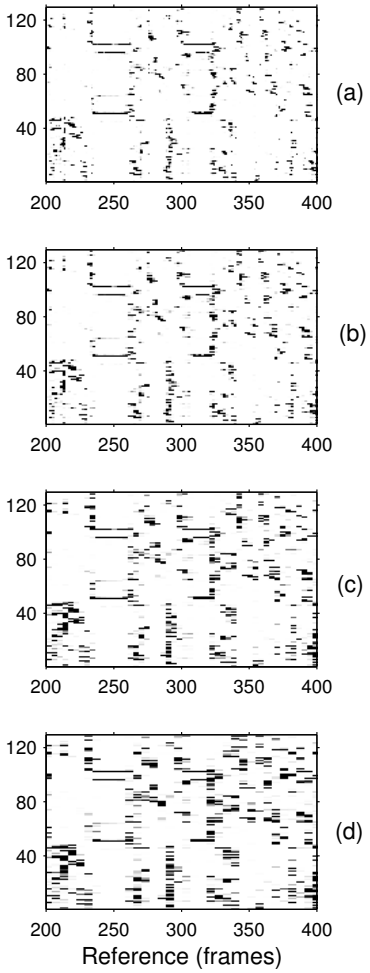


Fig. 10. Gaussian posteriorgrams of a reference segment for (a) $\alpha = 1$, (b) $\alpha = 2$, (c) $\alpha = 4$, (d) $\alpha = 6$. The y-axis represents the indices of the Gaussian components in GMM. Please note that the frames on the x-axis are repeated for α times to visualize the smoothed Gaussian posteriorgrams on the same scale. For visualization, we normalize each of the columns with the maximum value of the column.

Fig. 10 shows the 128 dimensional Gaussian posteriorgrams of a reference segment for $\alpha = 1, 2, 4, 6$, where $\alpha = 1$ represents no reduction in the Gaussian posteriorgrams. In Fig. 10, the frames on the x-axis are repeated for α times to visualize the smoothed Gaussian posteriorgrams on the same scale. From Fig. 10 and Table V, it is evident that for smaller values of α, such as $\alpha = 2$, the Gaussian posteriorgrams are similar to that of $\alpha = 1$ resulting in a fast search and yet obtaining a similar MTWV.

Fig. 11 show the alignment paths of FNS-DTW for $\alpha = 2, 4, 6$ (represented with dotted lines) in comparison with the alignment path of NS-DTW. The query and reference frames are reduced in FNS-DTW. For a graphical comparison with NS-DTW, the alignment path of FNS-DTW is stretched by a factor of α. From Fig. 11, it can be seen that the alignment path of FNS-DTW fluctuates around the alignment path of NS-DTW and the deviation is minimum for smaller values of α. This indicates that the *search hits* can be obtained by using FNS-DTW.

TABLE V
MISS PROBABILITY (MP), FALSE ALARM PROBABILITY (FAP) AND MAXIMUM TERM WEIGHTED VALUE (MTWV) OBTAINED FOR NS-DTW FOR VARIOUS VALUES OF α.

α	dev			eval		
	MP	FAP (10^{-2})	MTWV	MP	FAP (10^{-2})	MTWV
1	0.426	1.136	0.399	0.402	1.766	0.387
2	0.423	1.159	0.399	0.468	1.302	0.376
3	0.482	0.940	0.374	0.536	1.079	0.334
4	0.494	0.994	0.353	0.528	1.251	0.322
5	0.555	0.791	0.323	0.543	1.307	0.301
6	0.503	1.236	0.307	0.576	1.279	0.271

Table V shows the MTWV using FNS-DTW for dev and eval datasets for various values of α. The alignment path of FNS-DTW is similar to that of NS-DTW for smaller values of α. Thus the performance of FNS-DTW is much better for $\alpha = 2$ as compared to other values of α.

Fig. 12 shows QbE-STD runtime for FNS-DTW and NS-DTW (FNS-DTW for $\alpha = 1$). In Fast NS-DTW, there is a trade-off between search performance and accuracy. However, for low values of α ($\alpha = 2$) the MTWV is comparable to the original system on the dev dataset and slightly worse on the eval dataset (as shown in Table V). From Fig. 12 it is evident that FNS-DTW is 4 times faster than NS-DTW for $\alpha = 2$.

[19] describes a fast indexing based search approach called Randomized Acoustic Indexing and Logarithmic-time Search (RAILS) whose results were reported for MediaEval 2012 database. RAILS technique is as follows: (a) Locality sensitive hashing for indexing the data, (b) Approximate nearest neighbor search for each query frame in logarithmic time and constructing a similarity matrix, (c) Image processing techniques applied on the similarity matrix to obtain the *search hits*. The computation performance of the system was measured by the total size of the database in seconds divided by the average search time in seconds per query. The measure was referred to as *speedup*.

In [19], two search systems, RAILS-I and RAILS-II, were evaluated on MediaEval 2012 dev data and MTWV and *speedup* reported are shown in Table VI. From Table VI, it is shown the FNS-DTW-I (FNS-DTW for $\alpha = 2$) and FNS-DTW-II ($\alpha = 4$) are performing better than the RAILS system [19].

In [17], hierarchical K-Means clustering is used as an indexing technique and subsequently for computing the DTW scores. The *estimated speedup time* as reported on MediaEval 2012 dev data is 2400X with an MTWV of 0.364. In FNS-DTW with $\alpha = 4$, a *speedup* of 4100X is obtained with a slightly lower MTWV of 0.353 on the same dataset.

In other relevant works of [32], [33], a constraint based search was used to prune out the audio references. The

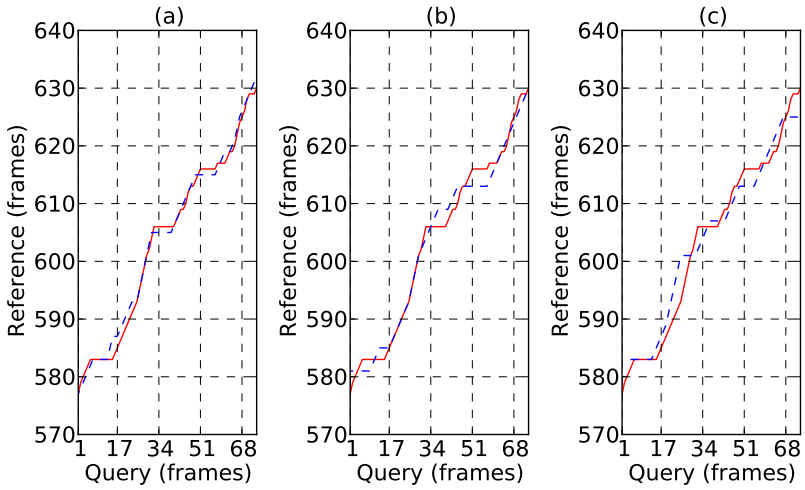


Fig. 11. Alignment paths for an example query and reference using NS-DTW and FNS-DTW using (a) $\alpha = 2$, (b) $\alpha = 4$, (c) $\alpha = 6$.

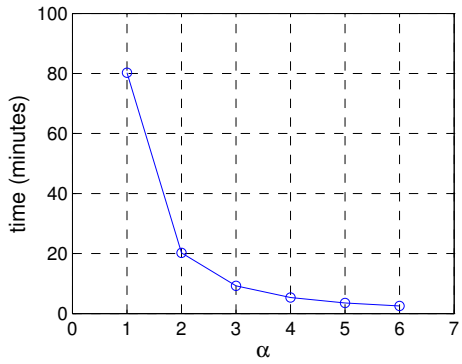


Fig. 12. Runtime of FNS-DTW for various $\alpha = 1, 2, 3, 4, 5, 6$ using dev dataset. This curve follows the trend of $\frac{1}{\alpha^2}$ due to computational upper bound of FNS-DTW being $O(\frac{mn}{\alpha^2})$.

TABLE VI
MTWV AND SPEEDUP FOR FNS-DTW AND RAILS EVALUATED ON DEV DATA.

	MTWV	**Speedup**
RAILS-I	0.381	1000X
FNS-DTW-I	0.399	1000X
RAILS-II	0.331	1600X
FNSDTW-II	0.353	4100X

pruning process was implemented by computing a lower bound estimate for DTW. It was shown that the computation of lower bound estimate is of the order $O(mn)$ [33]. Thus the total computational upper bound for such approaches would be $O(mn)$ plus the time taken to perform DTW alignment score. In our proposed fast NS-DTW, we use the reduced feature representation by averaging the successive Gaussian posteriorgrams. Thus the total computation time of fast NS-DTW would be $O(mn)$ plus the time taken to smooth the average the posteriorgrams. It should be noted that the fast NS-DTW is a one-stage process, whereas the lower bound estimate methods are implemented in two stages (pruning and score estimation).

VII. Conclusion and Future Work

In this paper we used a DTW based algorithm called non-segmental DTW (NS-DTW), with a computational upper bound of $O(mn)$. We have analyzed the performance of NS-DTW for query-by-example spoken term detection (QbE-STD) with Gaussian posteriorgrams obtained from different features of the speech signal. The results indicate that frequency domain linear prediction cepstral coefficients (FDLP), which capture the temporal dynamics of the speech signal, can be used as an alternative to traditional spectral features such as linear prediction cepstral coefficients (LPCC), perceptual linear prediction cepstral coefficients (PLP) and Mel-frequency cepstral coefficients (MFCC).

We have introduced a fast NS-DTW (FNS-DTW) which uses reduced Gaussian posteriorgrams for QbE-STD. We have shown that, for a given reduction factor $\alpha \in \mathbb{N}$, the computational upper bound of FNS-DTW is $O(\frac{mn}{\alpha^2})$. The reduction of the feature vectors was done via arithmetic mean and it was shown that for $\alpha = 2$, maximum term weighted values (MTWV) of FNS-DTW were similar or slightly lower to that of NS-DTW but three times faster.

We have also compared FNS-DTW with a fast indexing based search approach called Randomized Acoustic Indexing and Logarithmic-time Search (RAILS) whose results were reported for MediaEval 2012 database. It was shown that FNS-DTW was performing better than RAILS system with 0.353 MTWV search performance and a *speedup* of 4100X. One of the primary advantages of RAILS system over FNS-DTW is its indexing based technique to search over large databases and hence RAILS performance is better in terms of memory consumption. As a future work we plan to incorporate indexing based techniques in building sparse similarity matrix for FNS-DTW type of approach.

ACKNOWLEDGEMENTS

We would also like to thank Tata Consultancy Services (TCS) for partially supporting Gautam's PhD fellowship at International Institute of Information Technology - Hyderabad, India.

REFERENCES

[1] I. Szöke, M. Fapso, L. Burget, and J. Cernocky, "Hybrid word-subword decoding for spoken term detection," in *Workshop on Searching Spontaneous Conversational Speech*, 2008, pp. 4–11.
[2] M. Saraclar and R. Sproat, "Lattice-based search for spoken utterance retrieval," in *Proc. of HLT-NAACL*, 2004, pp. 129–136.
[3] D. R. H. Miller, M. Kleber, C.-L. Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, and H. Gish, "Rapid and accurate spoken term detection," in *Proc. of INTERSPEECH*, 2007, pp. 314–317.
[4] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *Proc. of ASRU*, 2009, pp. 398–403.
[5] C.-A. Chan and L.-S. Lee, "Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping," in *Proc. of INTERSPEECH*, 2010, pp. 693–696.
[6] V. Gupta, J. Ajmera, A., and A. Verma, "A language independent approach to audio search," in *Proc. of INTERSPEECH*, 2011, pp. 1125–1128.
[7] A. Jansen and B. V. Durme, "Efficient spoken term discovery using randomized algorithms," in *Proc. of ASRU*, 2011, pp. 401–406.
[8] X. Anguera, "Speaker independent discriminant feature extraction for acoustic pattern-matching," in *Proc. of ICASSP*, 2012, pp. 485–488.
[9] X. Anguera and M. Ferrarons, "Memory efficient subsequence DTW for query-by-example spoken term detection," in *Proc. of ICME*, 2013.
[10] K. Prahallad, A. R. Toth, and A. W. Black, "Automatic building of synthetic voices from large multi-paragraph speech databases," in *Proc. of INTERSPEECH*, 2007, pp. 2901–2904.
[11] K. Prahallad and A. W. Black, "Segmentation of monologues in audio books for building synthetic voices," *IEEE Trans. Audio, Speech and Lang. Processing*, vol. 19, no. 5, pp. 1444–1449, July 2011.
[12] M. Müller, *Information Retrieval for Music and Motion*, Springer, Inc., 2007.
[13] F. Metze, N. Rajput, X. Anguera, M. H. Davel, G. Gravier, C. J. V. Heerden, G. V. Mantena, A. Muscariello, K. Prahallad, I. Szöke, and J. Tejedor, "The spoken web search task at MediaEval 2011," in *Proc. of ICASSP*, 2012, pp. 5165–5168.
[14] S. Thomas, S. Ganapathy, and H. Hermansky, "Recognition of reverberant speech using frequency domain linear prediction," *IEEE Signal Processing Letters*, vol. 15, pp. 681 –684, 2008.
[15] S. Ganapathy, S. Thomas, and H. Hermansky, "Temporal envelope compensation for robust phoneme recognition using modulation spectrum," *Journal of Acoustical Society of America*, vol. 128, pp. 3769–3780, 2010.
[16] M. Athineos and D.P.W. Ellis, "Autoregressive modeling of temporal envelopes," *IEEE Trans. Signal Processing*, vol. 55, no. 11, pp. 5237–5245, Nov. 2007.
[17] G. Mantena and X. Anguera, "Speed improvements to information retrieval-based dynamic time warping using hierarchical k-means clustering," in *Proc. of ICASSP*, 2013.
[18] E. Barnard, M. H. Davel, and C. J. V. Heerden, "ASR corpus design for resource-scarce languages," in *Proc. of INTERSPEECH*, 2009, pp. 2847–2850.
[19] A. Jansen, B. V. Durme, and P. Clark, "The JHU-HLTCOE spoken web search system for MediaEval 2012," in *MediaEval*, 2012.
[20] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of Acoustical Society of America*, vol. 57, no. 4, pp. 1738–52, Apr. 1990.
[21] J. Makhoul, "Linear prediction: A tutorial review," *Proc. of IEEE*, vol. 63, no. 4, pp. 561 – 580, april 1975.
[22] S. Ganapathy, *Signal analysis using autoregressive models of amplitude modulation*, Ph.D. thesis, Johns Hopkins University, Baltimore, Maryland, USA, Jan. 2012.
[23] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 26, pp. 43–49, 1978.
[24] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 23, no. 1, pp. 67–72, Feb. 1975.
[25] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 1, pp. 23–31, 2005.
[26] H. Ney and A. Noll, "Phoneme modelling using continuous mixture densities," in *Proc. of ICASSP*, 1988, pp. 437–440.
[27] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddington, "Results of the 2006 spoken term detection evaluation," in *Proc. of Workshop on Searching Spontaneous Conversational Speech*, 2007, pp. 45–50.
[28] F. Metze, E. Barnard, M. H. Davel, C. J. V. Heerden, X. Anguera, G. Gravier, and N. Rajput, "The spoken web search task," in *MediaEval*, 2012.
[29] R. Singh, B. Lambert, and B. Raj, "The use of sense in unsupervised training of acoustic models for ASR systems," in *Proc. of INTERSPEECH*, 2010, pp. 2938–2941.
[30] A. Jansen and K. Church, "Towards unsupervised training of speaker independent acoustic models," in *Proc. of INTERSPEECH*, 2011, pp. 1693–1692.
[31] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, Oct. 2007.
[32] Y. Zhang and J. Glass, "An inner-product lower-bound estimate for dynamic time warping," in *Proc. of ICASSP*, 2011, pp. 5660–5663.
[33] P. Yang, L. Xie, Q. Luan, and W. Feng, "A tighter lower bound estimate for dynamic time warping," in *Proc. of ICASSP*, 2013.

Gautam Mantena (S'13) received the B.Tech. degree from Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, Gujarat, India, in 2006 and the M.S. in IT degree from International Institute of Information Technology, Hyderabad (IIIT-H), India, in 2008. He is currently pursuing the Ph.D. degree from Speech and Vision Lab, IIIT-H. His research interests include spoken audio search, spoken dialogue systems and speech recognition.

Sivanand Achanta received the B.Tech. degree in Electronics and Communication from Kamala Institute of Technology and Science, Karimnagar,India in 2010. He is currently pursuing the Ph.D. degree from Speech and Vision Lab, International Institute of Information Technology,Hyderabad. His research interests include speech signal processing, machine learning and speech synthesis.

Kishore Prahallad (M'07) received the B.E. degree from the Deccan College of Engineering and Technology, Osmania University, Hyderabad, India, in 1998, the M.S. (by Research) degree from the Indian Institute of Technology (IIT) Madras, in 2001 and the Ph.D. degree from the Language Technologies Institute, School of Computer Science, Carnegie Mellon University (CMU), Pittsburgh, USA, in 2010. He is an Associate Professor at the International Institute of Information Technology, Hyderabad (IIIT-H). He has been associated with IIIT-H since March 2001, and started the speech activities in Language Technologies Research Center at IIIT-H. His research interests are in speech and language processing, multimodal mobile computing and interfaces, and artificial neural networks.

USE OF ARTICULATORY BOTTLE-NECK FEATURES FOR QUERY-BY-EXAMPLE SPOKEN TERM DETECTION IN LOW RESOURCE SCENARIOS

Gautam Mantena, Kishore Prahallad

International Institute of Information Technology - Hyderabad, India

gautam.mantena@research.iiit.ac.in, kishore@iiit.ac.in

ABSTRACT

For query-by-example spoken term detection (QbE-STD), generation of phone posteriorgrams requires labelled data which would be difficult for languages with low resources. One solution is to build models from rich resource languages and use them in the low resource scenario. However, phone classes are not language universal and alternate representation such as articulatory classes is explored. In this paper, we use articulatory information and their derivatives such as bottle-neck (BN) features (also referred to as articulatory BN features) for QbE-STD. We obtain Gaussian posteriorgrams of articulatory BN features in tandem with the acoustic parameters such as frequency domain linear prediction cepstral coefficients to perform the search. We compare the search performance of articulatory and phone BN features and show that articulatory BN features are a better representation. We also provide experimental results to show that low amounts (30 mins) of training data could be used to derive articulatory BN features.

Index Terms— Query-by-example spoken term detection, multi-layer perceptron, articulatory features, bottle-neck features, low resource.

1. INTRODUCTION

The task of a query-by-example spoken term detection (QbE-STD) is to search a spoken query in a spoken audio data. A traditional QbE-STD approach is to convert spoken audio into a sequence of symbols and then perform text based search. In [1–3], the audio is first converted to a sequence of symbols using large vocabulary continuous speech recognition (LVCSR) and then lattice based search techniques are incorporated. LVCSR based approaches have been shown to be accurate for well resourced languages. However, such approaches are not scalable for languages where there is no availability or the resources to build an LVCSR system. To overcome this limitation dynamic time warping (DTW) based techniques are exploited for QbE-STD [4–8].

Phone [4, 5] and Gaussian posteriorgrams [6–8] are some of the feature representations used for DTW-based QbE-STD. Generation of phone posteriorgrams require labelled data which would be difficult for languages with low resources. One solution is to build models from rich resource languages and use them in the low resource scenario [5, 9]. However, phone classes are not language universal and thus alternate representation such as articulatory classes is explored. Articulatory classes are language independent representation of speech sounds and classifiers could be trained on relatively low amounts of data [10, 11]. Articulatory information has been extensively used in LVCSR for (a) Robust recognition in noisy conditions [11–13], and (b) Multi-lingual and cross-lingual speech recognition [14–16]. In [17], spoken audio is decoded to a sequence of articulatory classes which is used to prune out the spoken audio before performing the DTW-based search.

In this paper, we use articulatory information and their derivatives such as bottle-neck (BN) features (also referred to as articulatory BN features) for QbE-STD. BN features have been used extensively in multi-lingual LVCSR and were shown to improve the word error rate [18–21]. In the context of QbE-STD, BN features of phone classes have been used to build a hierarchical neural network structure (referred to as BN universal context network) [22]. To our knowledge, BN features of articulatory classes have not been explored in the context of DTW-based QbE-STD.

The contributions of our work are as follows: (a) Use of articulatory information and its derivatives such as BN features for QbE-STD, (b) Use of BN features in tandem with the acoustic parameters such as frequency domain linear prediction cepstral coefficients to compute Gaussian posteriorgrams, (c) Comparison of Gaussian posteriorgrams obtained using articulatory and phone BN features, and (d) Experimental results to show that low amounts of training data could be used to obtain articulatory BN features.

The organization of the paper is as follows: Section 2 describes the database used in this work. In Section 3, we describe the DTW-based algorithm used to perform the search. Section 4 describes the acoustic parameters of the speech signal and the computation of Gaussian posteriorgrams. Section 5 describes the use of articulatory BN features for QbE-STD and its comparison with the phone BN features. In Section 6, we provide experimental results to show that 20-30 mins of training data can be used to derive articulatory BN features.

2. DATABASE

The experiments conducted in this work use MediaEval 2012 data which is a subset of Lwazi database [23]. The data consists of audio recorded via telephone in 4 of 11 South African languages. We consider two data sets, development (dev) and evaluation (eval) which contain spoken audio (reference) and spoken query data. The statistics of the audio data is shown in Table 1.

Table 1: Statistics of MediaEval 2012 data.

Data	Utts	Total(mins)	Average(sec)
dev reference	1580	221.863	8.42
dev query	100	2.372	1.42
eval reference	1660	232.541	8.40
eval query	100	2.537	1.52

All the evaluations are performed using 2006 NIST evaluation criteria [24, 25] and the corresponding actual term weighted values (ATWV) and maximum term weighted values (MTWV) are re-

ported. To compute the ATWV and MTWV, an average miss probability and false alarm probabilities are computed for all the queries. In this paper, an optimum threshold to retrieve the search results is computed using the dev dataset. This threshold is then applied on the eval dataset to obtain the ATWV.

3. QBE-STD USING NON-SEGMENTAL DTW

QbE-STD is performed using a variant of DTW-based search referred to as non-segmental DTW (NS-DTW) [5, 8, 26]. Let $\mathcal{Q} = \{\mathbf{q_1}, \mathbf{q_2}, \ldots, \mathbf{q_i}, \ldots, \mathbf{q_n}\}$ be a spoken query (or query) containing n feature vectors. Let $\mathcal{R} = \{\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_j}, \ldots, \mathbf{u_m}\}$ be the spoken audio (or reference) containing m feature vectors.

Each of these feature vectors represent a Gaussian, articulatory or phone posteriorgrams as computed in Sections 4 and 5. The distance measure between a query vector $\mathbf{q_i}$ and a reference vector $\mathbf{u_j}$ is given by:

$$d(i,j) = -log\left(\frac{\mathbf{q_i}}{||\mathbf{q_i}||} \cdot \frac{\mathbf{u_j}}{||\mathbf{u_j}||}\right) \quad (1)$$

We define the term *search hit* as the region in the reference $\mathcal{R}$ that is likely to contain the query $\mathcal{Q}$. The query can start from any point in the reference. Initially, $S(1,j) = d(1,j)$, where $d(1,j)$ is the distance measure. The entries in the rest of the similarity matrix for NS-DTW is given by Eq. (2).

$$S(i,j) = min\left\{\begin{array}{c}\frac{d(i,j)+S(i-1,j-2)}{T(i-1,j-2)+1} \\ \frac{d(i,j)+S(i-1,j-1)}{T(i-1,j-1)+2} \\ \frac{d(i,j)+S(i-1,j)}{T(i-1,j)+1}\end{array}\right\}, \quad (2)$$

where T is called the transition matrix. $T(i,j)$ represents the number of transitions required to reach i,j from a start point. In order to detect the start and end time stamps of the *search hit*, we obtain the reference index that contains the best alignment score, i. e., the end point of the *search hit* is given by $j = \min_j\{S(n,j)\}$ for $j = 1, 2, ..., m$. Once the end point j is obtained, the corresponding start point could be obtained by a path trace back. Thus we obtain the location of the query in the reference.

4. FEATURE REPRESENTATION USING GAUSSIAN POSTERIORGRAMS

In general, Gaussian posteriorgrams are obtained by a two step process [7, 8]. In the first step, acoustic parameters such as Mel-frequency cepstral coefficients (MFCC) or frequency domain linear prediction cepstral coefficients (FDLP) are extracted from the speech signal. In the second step, Gaussian posteriorgrams are computed by training a Gaussian mixture model (GMM) on the speech data and the posterior probability obtained from each Gaussian is used to represent the acoustic parameter. In this paper, we train a GMM containing 128 Gaussians to obtain 128 dimensional Gaussian posteriorgrams.

In [8], we show that the Gaussian posteriorgrams of FDLP perform better than that of MFCC. In MFCC, the short-time spectral properties of the speech signal is captured. In order to capture the temporal dynamics of the speech signal, FDLP was developed [27–29].

A 25 ms window length with 10 ms shift is considered to extract 13 dimensional features along with delta and acceleration coefficients for MFCC and FDLP. An all-pole model of order 160 poles/sec and 37 filter banks are considered to extract FDLP. A set of 26 filter banks are used for computing MFCC.

Table 2: MTWV obtained using 128 dimensional Gaussian posteriorgrams (GPost.) of 39 dimensional MFCC and FDLP. The values indicated in the brackets show the ATWV computed for the eval dataset.

Feats.	dim.	GPost. dim.	MTWV (ATWV)	
			dev	eval
MFCC	39	128	0.377	0.325 (0.323)
FDLP	39	128	0.399	0.387 (0.358)

Table 2 shows the MTWV using 128 dimensional Gaussian posteriorgrams of 39 dimensional MFCC and FDLP. The search is performed using NS-DTW as described in Section 3. From Table 2, it can be seen that Gaussian posteriorgrams of FDLP performs better than that of MFCC. Hence, we are motivated to use FDLP as the acoustic features for QbE-STD. A more detailed analysis of the performance of NS-DTW using FDLP is described in [8].

To obtain Gaussian posteriorgrams of the acoustic parameters such as FDLP, no class information such as phone or articulatory classes is used. In this paper, we derive bottle-neck (BN) features from an articulatory model (also referred to as articulatory BN features). We show that the Gaussian posteriorgrams of articulatory BN features in tandem with FDLP perform better than that of FDLP. Section 5 describes the use of articulatory BN features in detail.

5. ARTICULATORY BOTTLE-NECK FEATURES

Availability of labelled data is an issue for building supervised models such as multi-layer perceptron (MLP). To overcome such an issue we train models on a high resource language and use it in a low resource scenario.

Table 3: Articulatory classes of speech sounds

Articulatory Property	Classes	# bits
Voicing	±voicing	1
Vowel length	short, long, diphthong	3
Vowel height	high, mid, low	3
Vowel frontness	front, central, back	3
Lip rounding	±rounding	1
Manner of articulation	stop, fricative, affricative nasal, approximant	5
Place of articulation	velar, alveolar, palatal, labial, dental	5
Aspiration	±aspiration	1
Silence	±silence	1

We train an articulatory MLP using 24 hours of labelled Telugu database consisting of 49 phones [30]. These 49 phones are represented by 23 articulatory classes which characterize the speech production process such as vowel properties, place of articulation, manner of articulation, etc. We modify the articulatory classes described in [31] to suit the training data available. We use nine different articulatory properties (as shown in Table 3). Each articulatory property is further divided into sub classes resulting in a 23 dimensional articulatory posteriorgram.

The architecture used for training an articulatory MLP is 39L 120N 13L 120N 23S. For comparison we also train a phone MLP with an architecture 39L 120N 13L 120N 49S. The integer values in the MLP architecture indicate the number of nodes, and L (linear), N (non-linear) and S (sigmoid) represent the activation functions in

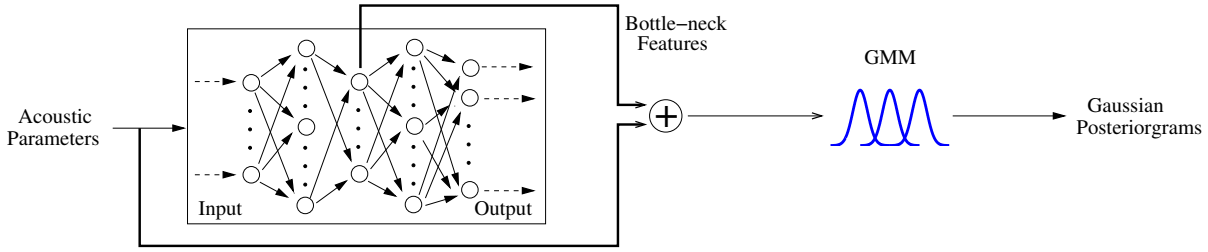


Fig. 1: A general block diagram for computing Gaussian posteriorgrams of bottle-neck features in tandem with the acoustic parameters such as FDLP.

each of the layers. We use 39 dimensional acoustic parameters as the input for the articulatory and phone MLPs.

Table 4 shows MTWV obtained using 23 dimensional articulatory, 49 dimensional phone and 128 dimensional Gaussian posteriorgrams. From Table 4, it can be seen that the Gaussian posteriorgrams perform better than the articulatory and phone posteriorgrams. Thus, phone and articulatory posteriorgrams under-perform when the language they were trained on differs from the target language [7, 32].

Table 4: MTWV obtained using 23 dimensional articulatory, 49 dimensional phone and 128 dimensional Gaussian posteriorgrams of FDLP. The values indicated in the brackets show the ATWV computed for the eval dataset.

Posteriorgrams	**Post. dim.**	**MTWV (ATWV)**	
		dev	eval
Art. Post.	23	0.212	0.172 (0.156)
Phone Post.	49	0.265	0.217 (0.209)
Gaussian Post.	128	**0.399**	**0.387 (0.358)**

5.1. Bottle-neck (BN) features

In order to exploit the class information captured by an MLP, we derive features from the bottle-neck layer (as shown in Fig. 1). These are referred to as bottle-neck (BN) features and are of 13 dimensions. The advantages of BN features are as follows [33]: (a) They are compressed features and are of lower dimension, and (b) Classification properties of the target class is reflected in the BN features.

5.2. Compressed (CP) features

An alternative representation to BN features can be obtained by post processing the articulatory posteriorgrams as follows: (a) A negative logarithm is applied on the articulatory posteriorgrams to scale the dynamic range and then followed by dimensionality reduction [14, 16]. These post processed posteriorgram features are referred to as compressed posteriorgram (CP) features, and (b) We then obtain Gaussian posteriorgrams of CP features in tandem with FDLP.

In the literature, CP features are referred to as tandem connectionist features [34] or probabilistic features [20, 33]. In [9], Gaussian posteriorgrams of CP features derived from phone MLPs were used for QbE-STD. However, it was shown that the Gaussian posteriorgrams of CP features were performing similar to that of the acoustic parameters. In this paper, we show that the search performance can be improved by using BN (or CP) features in tandem with the acoustic parameters such as FDLP.

To compress the log posteriorgram features, we perform a non-linear PCA using an auto associative neural network (AANN) with an architecture 23L 100N 13L 100N 23L. Thus we obtain 13 dimensional CP features from 23 dimensional articulatory posteriorgrams. These features are similar to that of the BN features as described in Section 5.1. However, an advantage of BN over CP features is that they do not require an explicit dimensionality reduction.

5.3. Comparison of BN and CP features

Table 5 shows MTWV obtained using Gaussian posteriorgrams of articulatory CP (AR-CP), articulatory BN (AR-BN), FDLP, FDLP + AR-CP and FDLP + AR-BN. From Table 5, it can be seen that: (a) Gaussian posteriorgrams of FDLP + AR-BN (or AR-CP) perform better than that of FDLP, and (b) Gaussian posteriorgrams of FDLP + AR-BN perform better than of FDLP + AR-CP. Thus we choose articulatory BN features to obtain Gaussian posteriorgrams for QbE-STD.

Table 5: MTWV obtained using Gaussian posteriorgrams of AR-CP, AR-BN, FDLP, FDLP + AR-CP and FDLP + AR-BN features. The values indicated in the brackets show the ATWV computed for the eval dataset.

Feats.	**dim.**	**GPost. dim.**	**MTWV (ATWV)**	
			dev	eval
AR-CP	13	128	0.336	0.331 (0.323)
AR-BN	13	128	0.419	0.390 (0.389)
FDLP	39	128	0.399	0.387 (0.358)
FDLP + AR-CP	52	128	0.465	0.467 (0.463)
FDLP + AR-BN	52	128	**0.494**	**0.492 (0.467)**

5.4. Selecting an Optimum Dimension for Articulatory BN Features

In this Section, we perform experiments to select an optimum dimension for AR-BN features. We derive AR-BN features of dimensions 5, 9, 13, 17 and 21 to obtain Gaussian posteriorgrams.

Fig. 2 shows MTWV obtained for dev data using Gaussian posteriorgrams of FDLP + AR-BN. We derive 5, 9, 13, 17 and 21 dimensional AR-BN features and use them in tandem with 39 dimensional FDLP parameters. MLP architecture used to derive AR-BN features is as follows: 23L 100N ΦL 100N 23L, where Φ = 5, 9, 13, 17, 21. From Fig. 2, it can be seen that the best performance is with 13 dimensional AR-BN features in tandem with FDLP. Thus, we choose 13 as the optimum AR-BN feature dimension.

5.5. Comparison with Phone BN Features

In this Section, we derive 13 dimensional phone BN features and compare it with articulatory BN features. The MLP architecture used to derive phone BN features is 39L 120N 13L 120N 49S. Table

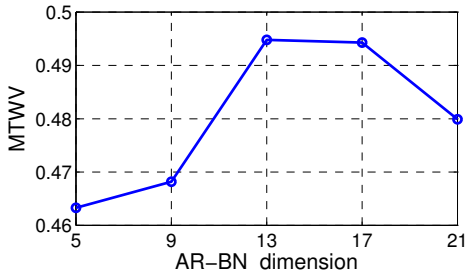


Fig. 2: MTWV obtained for dev data using Gaussian posteriorgrams of FDLP + AR-BN. AR-BN features of 5, 7, 13, 17 and 21 dimensions are used in tandem with 39 dimensional FDLP.

6 shows MTWV obtained using Gaussian posteriorgrams of phone and articulatory BN features in tandem with FDLP. The phone and articulatory BN features are denoted as PH-BN and AR-BN respectively.

Table 6: MTWV obtained using Gaussian posteriorgrams of FDLP + PH-BN and FDLP + AR-BN features. The values indicated in the brackets show the ATWV computed for the eval dataset.

Feats.	**dim.**	**GPost. dim.**	**MTWV (ATWV)**	
			dev	eval
FDLP + PH-BN	52	128	0.469	0.452 (0.425)
FDLP + AR-BN	52	128	**0.494**	**0.492 (0.467)**

From Table 6, it can be seen that the Gaussian posteriorgrams of FDLP + AR-BN perform better than that of FDLP + PH-BN. Articulatory classes are more language universal than phones. Thus AR-BN features are a better representation than PH-BN features to obtain Gaussian posteriorgrams.

6. USE OF LOW AMOUNTS OF TRAINING DATA FOR ARTICULATORY AND PHONE MLPS

In Section 5, we use an articulatory and phone MLPs trained on 24 hours of spoken audio data. However, access to such large amounts of labelled data is expensive and not always feasible. In this Section, we derive BN features from articulatory and phone MLPs trained on low amounts of spoken audio data.

Fig. 3 shows MTWV obtained for dev data using Gaussian posteriorgrams of FDLP, FDLP + PH-BN and FDLP + AR-BN. The articulatory and phone MLPs are trained using 10, 20, 30, 50 and 75 mins of audio data. MTWV obtained using Gaussian posteriorgrams of FDLP is the baseline performance and is denoted as an horizontal line (as shown in Fig. 3). From Fig. 3, we observe that 20-30 mins of training data can be used to derive AR-BN features. This is because each phone is represented by more than one articulatory class. This leads to a large amount of training material for each articulatory class, which often exceeds the amount of phone training data [11, 35].

Table 7 shows MTWV obtained using Gaussian posteriorgrams of FDLP, FDLP + PH-BN and FDLP + AR-BN. PH-BN and AR-BN features are derived from MLPs trained on 30 mins of labelled data. From Table 7, it can be seen that 30 mins can be used to derive AR-BN features to obtain Gaussian posteriorgrams. However, there is a trade-off between the performance of the BN features and the amount of data used for training (as shown in Fig. 3)

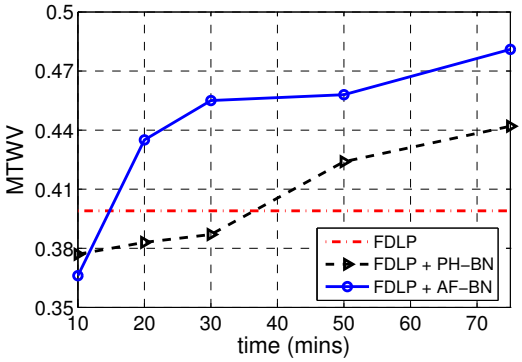


Fig. 3: MTWV obtained for dev data using Gaussian posteriorgrams of FDLP, FDLP + PH-BN and FDLP + AR-BN. The x-axis represent the amount of labelled data used to train the MLPs.

Table 7: MTWV obtained using Gaussian posteriorgrams of FDLP, FDLP + PH-BN and FDLP + AR-BN. The BN features are obtained from 30 mins of training data. The values indicated in the brackets show the ATWV computed for the eval dataset.

Feats.	**dim.**	**GPost. dim.**	**MTWV (ATWV)**	
			dev	eval
FDLP	39	128	0.399	0.387 (0.358)
FDLP + PH-BN	52	128	0.387	0.391 (0.338)
FDLP + AR-BN	52	128	**0.455**	**0.442 (0.425)**

7. CONCLUSIONS

In this paper, we have used articulatory information and its derivatives such as bottle-neck (BN) features (also referred to as articulatory BN features) for query-by-example spoken term detection (QbE-STD). We compared the search performance using Gaussian posteriorgrams of articulatory BN (AR-BN) and phone BN (PH-BN) features and have shown that AR-BN features are a better representation. We have also provided experimental results to show that 30 mins of training data could be used to derive AR-BN features.

Acknowledgements

We would like to thank Florian Metze, CMU for clarification on the literature of articulatory and bottle-neck features. We would also like to thank Tata Consultancy Services (TCS) for partially supporting Gautam's PhD fellowship at IIIT-H, India.

8. REFERENCES

[1] I. Szöke, M. Fapso, L. Burget, and J. Cernocky, "Hybrid word-subword decoding for spoken term detection," in *Workshop on Searching Spontaneous Conversational Speech*, 2008, pp. 4–11.

[2] M. Saraclar and R. Sproat, "Lattice-based search for spoken utterance retrieval," in *Proc. of HLT-NAACL*, 2004, pp. 129–136.

[3] D. R. H. Miller, M. Kleber, C.-L. Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, and H. Gish, "Rapid and accurate spoken term detection," in *Proc. of INTERSPEECH*, 2007, pp. 314–317.

[4] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Proc. of ASRU*, 2009, pp. 421–426.

[5] V. Gupta, J. Ajmera, A., and A. Verma, "A language independent approach to audio search," in *Proc. of INTERSPEECH*, 2011, pp. 1125–1128.

[6] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *Proc. of ASRU*, 2009, pp. 398–403.

[7] X. Anguera, "Speaker independent discriminant feature extraction for acoustic pattern-matching," in *Proc. of ICASSP*, 2012, pp. 485–488.

[8] G. Mantena, S. Achanta, and K. Prahallad, "Query-by-example spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping," *accepted for publication in IEEE Trans. Audio, Speech and Lang. Processing*, 2014.

[9] H. Wang, T. Lee, C.-C. Leung, B. Ma, and H. Li, "Using parallel tokenizers with DTW matrix combination for low-resource spoken term detection," in *Proc. of ICASSP*, 2013.

[10] A. W. Black, T. Bunnell, Y. Dou, P. Muthukumar, F. Metze, D. Perry, T. Polzehl, K. Prahallad, S. Steidl, and C. Vaughn, "Articulatory features for expressive speech synthesis," in *Proc. of ICASSP*, Kyoto, Japan, 2012.

[11] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, no. 3-4, pp. 303–319, 2002.

[12] K. Livescu, Ö. Cetin, M. Hasegawa-johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, A. Bezman, S. Dawson-Haggerty, B. Woods, J. Frankel, M. Magimai-Doss, and K. Saenko, "Audiovisual speech recognition with articulator positions as hidden variables," in *in Proc. of ICASSP*, 2007.

[13] V. Mitra, W. Wang, A. Stolcke, H. Nam, C. Richey, J. Yuan, and Mark Liberman, "Articulatory features for large vocabulary speech recognition," in *in Proc. of ICASSP*, 2013.

[14] Ö. Çetin, M. Magimai-Doss, K. Livescu, A. Kantor, S. King, C. D. Bartels, and J. Frankel, "Monolingual and crosslingual comparison of tandem features derived from articulatory and phone MLPs," in *Proc. of ASRU*, 2007, pp. 36–41.

[15] S. Stüker, F. Metze, T. Schultz, and A. Waibel, "Integrating multilingual articulatory features into speech recognition.," in *in Proc. INTERSPEECH*, 2003.

[16] L. Tóth, J. Frankel, G. Gosztolya, and S. King, "Cross-lingual portability of MLP-based tandem features - a case study for English and Hungarian," in *Proc. of INTERSPEECH*, 2008, pp. 2695–2698.

[17] F. Metze, N. Rajput, X. Anguera, M. H. Davel, G. Gravier, C. J. V. Heerden, G. V. Mantena, A. Muscariello, K. Prahallad, I. Szöke, and J. Tejedor, "The spoken web search task at MediaEval 2011," in *Proc. of ICASSP*, 2012, pp. 5165–5168.

[18] F. Grézl and P. Fousek, "Optimizing bottle-neck features for LVCSR," in *Proc. of ICASSP*, 2008, pp. 4729–4732.

[19] N. T. Vu, F. Metze, and T. Schultz, "Multilingual bottle-neck features and its application for under-resourced languages," in *Proc. of SLTU*, 2012.

[20] F. Grézl, M. Karafiát, and M. Janda, "Study of probabilistic and bottle-neck features in multilingual environment," in *Proc. of ASRU*, 2011, pp. 359–364.

[21] K. Vesely, M. Karafiat, F. Grezl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Proc. of SLT*, 2012, pp. 336–341.

[22] J. Tejedor, I. Szöke, and M. Fapso, "Novel methods for query selection and query combination in query-by-example spoken term detection," in *in Proc. of SSCS*, 2010, pp. 15–20.

[23] E. Barnard, M. H. Davel, and C. J. V. Heerden, "ASR corpus design for resource-scarce languages," in *Proc. of INTERSPEECH*, 2009, pp. 2847–2850.

[24] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddington, "Results of the 2006 spoken term detection evaluation," in *Proc. of Workshop on Searching Spontaneous Conversational Speech*, 2007, pp. 45–50.

[25] F. Metze, E. Barnard, M. H. Davel, C. J. V. Heerden, X. Anguera, G. Gravier, and N. Rajput, "The spoken web search task," in *MediaEval*, 2012.

[26] X. Anguera and M. Ferrarons, "Memory efficient subsequence DTW for query-by-example spoken term detection," in *Proc. of ICME*, 2013.

[27] S. Ganapathy, *Signal analysis using autoregressive models of amplitude modulation*, Ph.D. thesis, Johns Hopkins University, Baltimore, Maryland, USA, Jan. 2012.

[28] S. Thomas, S. Ganapathy, and H. Hermansky, "Recognition of reverberant speech using frequency domain linear prediction," *IEEE Signal Processing Letters*, vol. 15, pp. 681 –684, 2008.

[29] S. Ganapathy, S. Thomas, and H. Hermansky, "Temporal envelope compensation for robust phoneme recognition using modulation spectrum," *Journal of Acoustical Society of America*, vol. 128, pp. 3769–3780, 2010.

[30] G. K. Anumanchipalli, R. Chitturi, S. Joshi, S. Singh R. Kumar, R.N.V Sitaram, and S.P. Kishore, "Development of Indian language speech databases for LVCSR," in *Proc. of SPECOM*, Patras, Greece, 2005.

[31] B. Bollepalli, A. W. Black, and K. Prahallad, "Modelling a noisy-channel for voice conversion using articulatory features," in *Proc. of INTERSPEECH*, 2012.

[32] A. Muscariello, G. Gravier, and F. Bimbot, "Audio keyword extraction by unsupervised word discovery," in *Proc. of INTERSPEECH*, 2009, pp. 2843–2846.

[33] F. Grézl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. of ICASSP*, 2007, vol. 4, pp. 757–760.

[34] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. of ICASSP*, 2000, pp. 1635–1638.

[35] K. Kirchhoff, *Robust speech recognition using articulatory information*, Ph.D. thesis, Bielefeld University, 1999.

UNSUPERVISED QUERY-BY-EXAMPLE SPOKEN TERM DETECTION USING SEGMENT-BASED BAG OF ACOUSTIC WORDS

Basil George and B. Yegnanarayana

Speech and Vision Lab, International Institute of Information Technology, Hyderabad, India

basil.george@research.iiit.ac.in and yegna@iiit.ac.in

ABSTRACT

In this work, we present an unsupervised framework to address the problem of spotting spoken terms in large speech databases. The segment-based Bag of Acoustic Words (BoAW) framework proposed is inspired from the Bag of Words (BoW) approach widely used in text retrieval systems. Since this model ignores the sequence information in speech samples for efficient indexing of the database, a Dynamic Time Warping (DTW) based temporal matching technique is used to re-rank the results and restore the time sequence information. The speech data is stored efficiently in an inverted index which makes the retrieval very fast, thus making this framework particularly useful for searching large databases. We address the issue of choosing the appropriate size of the segment of speech for reliable indexing. Comparison with other query-by-example spoken term detection systems shows that the proposed system outperforms the rest.

Index Terms— query-by-example, spoken term detection, Bag of Acoustic Words, template matching, unsupervised learning, segment ranking

1. INTRODUCTION

In the digital era, huge amount of audio data is being produced and consumed every day in a large variety of languages. This may be in the form of music, TV news, classroom lectures, audio books, podcasts, call center archives and even personal audio recordings. With this exponential growth of digital multimedia content, audio search becomes essential for fast retrieval of information from audio archives. Query-by-example (QbE) spoken term detection (STD) is a speech search framework in which spoken queries are used to retrieve matching portions from a speech database.

State of the art approaches rely on automatic speech recognition (ASR) frameworks which have shown good performance in well-resourced contexts [1, 2]. But, such LVCSR-based systems can only be built for resource-rich languages where huge amounts of transcribed speech data is available to train statistical and acoustical models. Another requirement for good performance of ASR based systems is the large vocabulary coverage during the training phase so that out-of-vocabulary (OOV) terms are not presented for recognition during the searching phase. This may not be possible in practical systems, thus causing higher word error rates (WER) and deteriorating the overall performance. Though some methods to tackle the OOV problem like making the system vocabulary independent, sub-word unit modeling of OOV terms, phonetic search frameworks etc. have been proposed, it continues to be a challenging task [3, 4, 5, 6].

2. RELATION TO PRIOR WORK

Due to various limitations of ASR-based systems, template matching based methods for QbE STD have been explored in recent years [7, 8, 9, 10, 11]. In these methods, audio data is stored as templates that are generated by acoustic-phonetic models. When a spoken query is presented to the system, its template is generated, which is then searched in the database, typically by using a variant of the Dynamic Time Warping (DTW) algorithm. Recently, the posteriorgram representation has become a very popular choice for the template [7, 8, 10, 12]. It is a representation of speech as a sequence of posterior probability vectors. Each vector denotes the posterior probability of a speech frame belonging to different classes. Depending on the way these classes are defined, different posteriorgrams such as phonetic, neural-network and Gaussian posteriorgrams are obtained.

But the absence of efficient indexing techniques makes posteriorgram-based systems not scalable for practical use, as the entire database is searched in a linear fashion even for very short queries. Recently, some attempts have been made to address this limitation by using locality sensitive hashes and subspace-indexing techniques for efficient storage of speech data [13, 14]. In this work, we propose an inverted indexing framework using Gaussian posteriorgrams for achieving fast reduction of the search space. The segment-based Bag of Acoustic Words (BoAW) framework proposed is inspired from the Bag of Words (BoW) model widely used in text retrieval systems. In recent years, similar techniques have been explored in other related fields such as object matching in videos, word image retrieval etc. which have shown great potential [15, 16].

3. BAG OF ACOUSTIC WORDS AND INVERTED INDEX

The BoAW model used in this work is inspired from the Bag of Words (BoW) model widely employed in text retrieval systems. A spoken document can be represented as an unordered collection of discrete acoustic units. These discrete acoustic units are termed as acoustic words. The acoustic words may be interpreted as the sounds or the frame-wise phonetic content present in the documents. Each document gets represented as a bag of discrete acoustic words. Similarly, a spoken document can be represented as a bag of syllables or a bag of spoken lexical words. The challenge in these approaches is to reliably segment speech into syllables or words. The work presented in this paper can be described as a bag of discrete sounds in which the frame-wise phonetic information of speech is chosen as the acoustic unit of the BoAW model. In this work, a GMM-based soft clustering approach is used which models the speech using a set of Gaussian distributions. The number of such distributions (K) is predetermined and can be loosely associated with the number of phonetic units present in the data.

The K mean vectors and covariance matrices obtained after this unsupervised training phase becomes the vocabulary of the system. This audio vocabulary is then used to quantize the extracted features by choosing the clusters with the highest posterior probabilities. The final representation for a spoken document is the frequency counts or a histogram of the quantized acoustic features $[f_1, f_2, ...f_i, ...f_K]$, where f_i is the number of occurrences of the i^{th} cluster or acoustic word in the spoken document and K is the vocabulary size. The differences in the durations of different spoken documents is accounted for by normalizing the BoAW histogram with respect to the segment size. From this normalized histogram, those acoustic words or clusters having frequency above a threshold (δ) are chosen to represent the document in the inverted index. These are termed as *'significant acoustic words'* of the document. The inverted index is an indexed data structure which stores a mapping from content to locations in the database. The location of the document in the database is associated with the significant acoustic words in the inverted index. Once the entire database is indexed, the location of every spoken document can be determined from the significant acoustic words obtained from that document.

An important issue to note in this approach is the loss of temporal information of speech. For example, the words 'tale' and 'late' may have the same phonetic content and hence, similar histogram representations, which reduces the precision of the system during the retrieval task. We address this issue, while exploiting the computational advantages of the BoAW approach, as explained in the next section. Another crucial point is the duration of a spoken document that should go into the index. The duration of the segments should be chosen in such a way that the significant acoustic words obtained from their BoAW histogram should be able to fully represent these segments. The segment histogram must be robust enough to reduce the false positive and false negative rates while maintaining the time taken for retrieval within practical terms. Detailed experiments are conducted to obtain the optimum segment size to index the documents for queries of different durations. In this segment-based inverted indexing paradigm, the term 'spoken document' will now be referred to as a 'segment' as it is a segment of speech, along with its time information (location within a file), that goes into the index. Figure 1 shows the Gaussian posteriorgram and BoAW histogram of a segment of length $1s$.

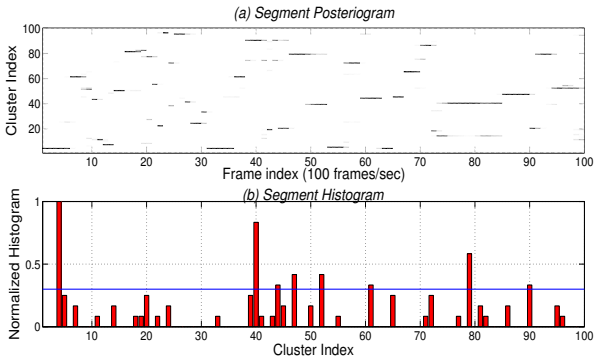


Fig. 1. (a) Gaussian posteriorgram and (b) normalized BoAW histogram of a speech segment with K=100, along with histogram threshold $\delta = 0.3$ marked in the figure.

4. RETRIEVAL SYSTEM

The task of the retrieval system is to return the best matches of an audio query from the indexed database. The frame-wise features are extracted from the query and the BoAW histogram is generated using GMM clustering. The significant acoustic words from the histogram are obtained by using a threshold (δ_q), which may be different from the threshold (δ) used while indexing the database. The choice of the threshold needs to be determined experimentally to balance the false rejection rate, false acceptance rate and the amount of the search space reduction achieved. Using the significant acoustic words obtained from the query, the list of database segments associated with them are retrieved from the inverted index. This is a very quick process which helps in locating the most probable segments in the database which match with the query.

In the BoAW approach, the sequence information in speech was ignored while performing efficient database indexing. But, as was mentioned earlier, this reduces the effectiveness of the system due to the possibility of a large number of false acceptances. Hence, a Dynamic Time Warping (DTW) approach is used to restore the sequence information

in the retrieved segments. DTW is performed between the Gaussian posteriorgrams of the query segment and that of the most probable database segments. The distance function used for DTW is:

$$D(p, q) = -log(p.q) \quad (1)$$

where p and q are two Gaussian posterior vectors. The dot product gives the probability of these two vectors drawing from the same distribution [7].

The ranking of database segments is performed not only using the DTW score but also including the BoAW histogram score to form a final merged score. The histogram score of an indexed database segment is the number of times that segment is retrieved by the significant acoustic words of the query segment. For example, suppose a query segment has 50 significant acoustic words for a vocabulary size of 100. For each of these 50 words, the system retrieves the most probable database segments from the inverted index. Suppose a particular indexed segment s_i was retrieved by 40 of these significant acoustic words. Then, the histogram score of the segment s_i is 40. Higher the histogram score, higher the probability of the segment being a correct match. The merged score S_{M_i} of a database segment s_i is computed as:

$$S_{M_i} = \alpha.S_{DTW_i} + \frac{\beta}{S_{Hist_i}} \quad (2)$$

where S_{DTW_i} and S_{Hist_i} are the DTW and histogram scores of the segment s_i, respectively, and α and β are scaling parameters which are determined empirically. Lower values of the DTW and merged scores are expected from matching portions. Thus, the database segments are ranked in the ascending order of their merged scores, and are presented as the output of the system (file name and time stamp).

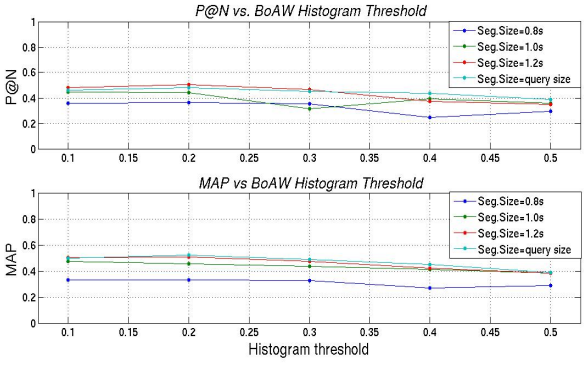


Fig. 2. (a) P@N and (b) MAP scores vs. BoAW histogram threshold δ with δ_q=δ, α=0.8, β=10(1-α), γ_1=1, γ_2=400 and Q=10.

5. EXPERIMENTS AND EVALUATION

This unsupervised QbE STD framework is tested on the TIMIT corpus using 30 queries of varying lengths. The TIMIT corpus is divided into 3 sets: development set (1000 files, 50 minutes), database set (4500 files, 3.8 hours) and test set (800 files). The development set is used to obtain the vocabulary using unsupervised GMM training of frame-wise 39-dimensional MFCC features as explained in the previous sections. Once the GMM is trained for K clusters, the database set is divided into segments which are added to the inverted index. Queries presented to the system are excised from the test set utterances. The generalizing capability of this framework is evaluated by keeping all the three sets non-overlapping. A query has, on average, about 5 relevant occurrences in the database. Hence, the evaluation metrics used are: i) **P@1**: Average precision of the top result returned by the system; ii) **P@3**: Average precision of the top 3 results; iii) **P@5**: Average precision of the top 5 results; iv) **P@*N***: Average precision of the top N results, where N is the number of occurrences of each query in the database; v) **MAP**: Mean average precision which is the mean of the precision scores after each query hit is retrieved.

Table 1. *Precision scores for different segment sizes with threshold δ=0.2.*

	P@1	**P@3**	**P@5**	**P@*N***	**MAP**
Seg.Size = 0.8s	0.5357	0.4405	0.3643	0.3642	0.3294
Seg.Size = 1.0s	0.7333	0.5667	0.4400	0.4405	0.4570
Seg.Size = 1.2s	0.7333	0.6333	0.5200	**0.5028**	**0.5051**
Query-guided	0.8000	0.6222	0.4933	0.4789	0.5214

As mentioned earlier, the choice of the segment size becomes crucial in the overall performance of the system. To determine the optimum segment length, we conduct experiments with two kinds of segmentation: query-guided segmentation and hard segmentation. In hard segmentation, the entire database is indexed prior to query submission by dividing it into segments of a pre-determined duration. In this case, the query may also need to be segmented as its length may be much larger than the segment duration, which may result in highly skewed warping paths during the DTW. This segmentation of a query implies that the database is searched for those segments which match with each of the query segments. Hence, the system returns scored results pertaining to different segments of the query and not for the entire query altogether. Hence, we need a way of merging the nearby database segments and obtain a combined score for these portions using their individual segment scores. A novel scoring strategy is employed which uses the positional weights (w) and merged scores (S_M) of individual database segments to obtain the final scores. Each database segment, ranked in the ascending order of the merged score (S_M), is grouped with its

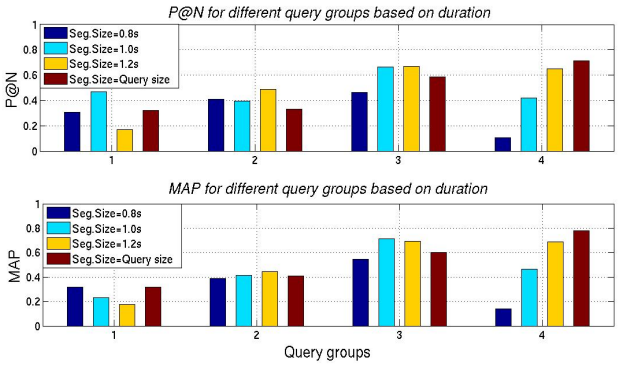


Fig. 3. (a) P@N and (b) MAP scores for different query groups (G) based on duration D_G. $0 < D_{G_1} < 0.8s$, $0.8s \leq D_{G_2} < 1.0s$, $1.0s \leq D_{G_3} < 1.2s$ and $1.2s \leq D_{G_4} < 1.6s$.

$(L-1)$ neighboring segments to form larger segments called files. The number of segments (L) in each file is fixed to match the query length. Suppose N such files are present in the database. S_{F_i} is the score of file i, taking into account the positions p_{ij} of the L segments within a file along with their merged scores $S_{M_{ij}}, j = 1, 2, ..L$, which is computed as:

$$S_{F_i} = \gamma_1.(w_{i1} + S_{M_{i1}}) - \gamma_2 \sum_{j=2}^{L} \frac{1}{w_{ij} S_{M_{ij}}} \quad (3)$$

where $i = 1, 2, ...N$ and

$$w_{ij} = \lfloor \frac{p_{ij} - 1}{Q} \rfloor + 1; j = 1, 2, ...L \quad (4)$$

where w_{ij} is the positional weight of the j^{th} segment of file i. Segment index j within a file is obtained by ranking the file segments using their merged scores. The scaling factors γ_1 and γ_2 are determined empirically. The quantization factor Q is used to divide the positional weights, depending on the initial ranking based on merged scores, into discrete levels. For example, results 1 through 10 and 11 through 20 are grouped into different levels, if $Q = 10$. This scoring criteria, given in (3), penalizes segments within a file based on their positional proximity and signal alignment to the best matched segment within a file. Such positional weighting, when combined with signal similarity scores, gives a good mechanism to rank different database files.

To compare the performance of the hard segmentation technique to a scenario where segmentation could be performed after a query is submitted, database is divided into overlapping segments of duration same as that of the query and populated into the inverted index. The histogram and DTW scores are merged and the segments are ranked. This experiment is conducted to study the correlation between database segment duration and query length.

6. RESULTS AND DISCUSSION

Figure 2 shows the relationship between P@N and MAP scores and histogram threshold (δ) with vocabulary size (K) as 100, query histogram threshold δ_q=δ and empirically determined scaling factors α=0.8, β=10(1-α), γ_1=1, γ_2=400. α is fixed to give greater weightage to the DTW score as compared to the histogram score. The quantization factor Q is set as 10. From the figure, we observe that the precision scores are maximum for $\delta = 0.2$. Table 1 gives precision scores for different segment sizes when $\delta = 0.2$. For a segment size of $1.2s$, P@N and MAP of 0.5028 and 0.5051, respectively, are obtained, which outperforms other systems proposed in literature. Table 2 shows the comparison of the proposed system with a system which uses a segmental variation of DTW [8] which is considered as the baseline for our experiments. To better understand the relationship between query duration and segment size, queries are grouped into groups (G) based of their duration (D_G). The durations of the groups are: $0 < D_{G_1} < 0.8s$, $0.8s \leq D_{G_2} < 1.0s$, $1.0s \leq D_{G_3} < 1.2s$ and $1.2s \leq D_{G_4} < 1.6s$. From figure 3, we see that precision scores are high when the query duration is large (groups 1 and 2). Also, for larger durational queries, segment size nearer to query size gives better results. This suggests that BoAW histogram representation becomes more reliable when a greater number of acoustic words are present in a segment. But segment size cannot be very different from the query size as it may lead to highly skewed warping paths. Hence, the segment size and the histogram threshold need to be chosen carefully to obtain the best results from the system.

Table 2. *Comparison of performance*

System	P@N
SDTW (#Examples=1)	0.4133
BoAW+DTW (proposed)	**0.5028**

7. CONCLUSION

In this paper, a new unsupervised framework for performing query-by-example spoken term detection was proposed. The Bag of Acoustic Words (BoAW) model enables efficient storage of speech in an inverted index data structure which helps in fast retrieval of matching segments. Further, temporal similarity is obtained by employing the Dynamic Time Warping technique. A new method of ranking audio documents which combines positional weights and similarity scores was also proposed. It was observed that the system gives very good performance when the query size is larger. In future, better segmentation techniques, such as those based on similarity of neighboring speech frames, need to be explored to help store the speech more efficiently.

8. REFERENCES

[1] David R. H. Miller, Michael Kleber, Chia-Lin Kao, Owen Kimball, Thomas Colthurst, Stephen A. Lowe, Richard M. Schwartz, and Herbert Gish, "Rapid and accurate spoken term detection," in *INTERSPEECH*, 2007, pp. 314–317.

[2] Murat Saraclar and Richard Sproat, "Lattice-based search for spoken utterance retrieval," in *HLT-NAACL*, 2004, pp. 129–136.

[3] Jonathan Mamou, Bhuvana Ramabhadran, and Olivier Siohan, "Vocabulary independent spoken term detection," in *SIGIR*, 2007, pp. 615–622.

[4] Igor Szöke, Lukás Burget, Jan Cernocký, and Michal Fapso, "Sub-word modeling of out of vocabulary words in spoken term detection," in *SLT*, 2008, pp. 273–276.

[5] Carolina Parada, Abhinav Sethy, and Bhuvana Ramabhadran, "Query-by-example spoken term detection for oov terms," in *ASRU*, 2009, pp. 404–409.

[6] Kenney Ng, *Subword-based approaches for spoken document retrieval*, Ph.D. thesis, Massachusetts Institute of Technology, 2000.

[7] T.J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *ASRU*, 2009, pp. 421–426.

[8] Yaodong Zhang and James R. Glass, "Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams," in *ASRU*, 2009, pp. 398–403.

[9] Chun an Chan and Lin-Shan Lee, "Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping," in *INTERSPEECH*, 2010, pp. 693–696.

[10] Vikram Gupta, Jitendra Ajmera, Arun Kumar, and Ashish Verma, "A language independent approach to audio search," in *INTERSPEECH*, 2011, pp. 1125–1128.

[11] Armando Muscariello, Guillaume Gravier, and Frédéric Bimbot, "Zero-resource audio-only spoken term detection based on a combination of template matching techniques," in *INTERSPEECH*, 2011, pp. 921–924.

[12] Guillermo Aradilla, Hervé Bourlard, and Mathew Magimai-Doss, "Posterior features applied to speech recognition tasks with user-defined vocabulary," in *ICASSP*, 2009, pp. 3809–3812.

[13] Aren Jansen and Benjamin Van Durme, "Indexing raw acoustic features for scalable zero resource search," in *INTERSPEECH*, 2012.

[14] Taisuke Kaneko and Tomoyosi Akiba, "Metric subspace indexing for fast spoken term detection," in *INTERSPEECH*, 2010, pp. 689–692.

[15] Josef Sivic and Andrew Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*, 2003, pp. 1470–1477.

[16] Ravi Shekhar and C. V. Jawahar, "Word image retrieval using bag of visual words," in *Document Analysis Systems*, 2012, pp. 297–301.

Acoustic Segmentation of Speech Using Zero Time Liftering (ZTL)

RaviShankar Prasad, B. Yegnanarayana

Speech and Vision Labs, IIIT Hyderabad

ravishankar.prasad@research.iiit.ac.in, yegna@iiit.ac.in

Abstract

Automatic segmentation of speech signals has been a constant engineering challenge. Even after the advances with supervised and unsupervised techniques, there still lies a challenge to equal the manually labelled segments. HMM-based segmentation techniques with modifications and corrections have been the state-of-art. These techniques are supervised in nature and thus require availability of large corpus transcribed with phone boundaries. The unsupervised techniques, on the other hand, explore gradients in various spectral and temporal properties of the speech signals. This paper presents a new and unsupervised method based on signal processing techniques to segment the speech signals. A recently developed method known as Zero Time Liftering (ZTL) is used for the analysis of speech signals, which provides fine temporal resolution of the spectral features of the segment being analyzed. It uses the Hilbert envelope of Numerator Group Delay (HNGD) of the signal to highlight its spectral activity. This representation is used to extract high SNR regions of the spectra, which in turn proves to be useful in representation of the production characteristics of the speech signal.Performance of the proposed analysis is at par with the existing baseline systems for unsupervised segmentation.

Index Terms: zero time liftering (ZTL), hilbert envelope of numerator group delay (HNGD), zero-frequency filter (ZFF), speech segmentation

1. Introduction

Segmentation of speech signals is a critical pre-processing task for several speech applications. These applications mostly rely on the availability of a corpus containing speech segments usually rich in linguistic and signal contents. Along with this, the corpus is also expected to contain information about its acoustic content with corresponding segment boundaries. The most precise way of maintaining such a corpus with well defined boundaries of speech and corresponding linguistic units with proper time alignment relies on manual efforts, till date. Manual segmentation and annotation of these databases require a large amount of time and effort, and therefore is a tedious job. Most speech processing applications generally use Hidden Markov Models (HMMs) and utilize the availability of such a corpus for training the models. Automatic speech segmentation techniques find their application with speech recognition, phonetic analysis, speech coding and other related areas of speech technology. A major advantage with automatic speech segmentation techniques is the consistency in their results. Studies indicate that manual labelling and segmentation processes are subjective, and thus may result in significant differences in the transcriptions by different people [1].

Automatic segmentation of speech signals is broadly classified into the explicit or implicit categories [2]. Explicit, being the text-dependent case, where a known phonetic sequence is time aligned against speech segments using a set of phone models or reference patterns. Implicit or text-independent techniques are those where there is no prior knowledge of corresponding phonetic sequence. Therefore, for these techniques, given a continuous speech, there is always non-compliance of the number of phones segments detected, to those actually present. In either of these cases the most commonly used approach includes trained HMM models. This process requires extraction of spectral characteristics of the speech signal followed by the forced alignment of HMM phone or syllable models using Viterbi alignment techniques. These techniques employing phone models to segment are termed as supervised techniques. In earlier attempts, a phone segmentation score of 87% was obtained using HMMs on TIMIT database [3]. Since then, several post-processing techniques have been combined with the HMM-based segmentation techniques to improve the detection scores for boundaries of phonetic segments. A statistical correction method was introduced to attune for the errors encountered with the HMM-based segmentation methods [4]. A speaker adaptation technique was also employed for error minimization along with the corrections of hypothesized boundaries to improve the results by almost 10% in both context-dependent and context-independent HMMs. Fusion techniques with multiple features [5] or multiple base segmentation engines [6] using regression methods have also been attempted to improve segmentation results.

The other class of techniques for automatic speech segmentation focuses on identifying changes in the signals in the temporal as well as in the spectral domain. These techniques emphasize on parameterization of speech signal and observing the behavior of these parameters over the entire signal. These are termed as unsupervised techniques and do not require any pre-acquired knowledge with respect to the data. These methods are bottom up approaches, where the lexical context integration is performed after the acoustic processing. In an attempt towards segmentation, the acoustic-phonetic knowledge of various manners and places of articulations was successfully employed with statistical pattern recognition approaches to obtain results comparable to HMM-based methods [7]. A comparative study, on phoneme segment detection using acoustic changes, with manually transcribed boundaries proved that unsupervised techniques are fairly effective and can be improved with infusion of additional information about broader classes based on energy, duration and articulatory cues [8]. An analysis of the hypothesized and missed boundaries with an unsupervised algorithm called the maximum margin clustering (MMC) was performed to improve the results obtained by the algorithm[9]. Another work proposed a delta spectral function (DSF) to represent the gradients in band energy for a specific band to measure the spectral changes [10]. Characterization of the rate of spectral transition to detect phoneme boundaries has also been employed to identify phoneme segment boundaries [11]. All these automatic

segmentation techniques have produced more or less similar results for a given tolerance level. The segmentation results are presented using the correct detection rate *(CDR)* $\boldsymbol{\alpha}$, which is given by,

$$\alpha(in\%) = \frac{N_{detected}}{N_{truth}} * 100 \quad (1)$$

where $N_{detected}$ refers to the total number of boundaries detected for a given tolerance level and N_{truth} is the number of boundaries in the ground truth segmentation. The tolerance level is generally expressed in *milliseconds*. For a segment boundary to be correct it has to fall within a tolerance window from the boundary marked for the ground truth.

In this paper we propose a method for unsupervised segmentation of speech based on recently developed signal processing techniques. In this method, information about the segments present or any related phone models is not used. The proposed method is independent of language, speaker or context. It has the advantage in terms of representation of vocal tract characteristics of the speech signal. The idea of the zero time liftering (ZTL) [12] method was conceived from a recently developed technique for identifying the glottal closure instants, the zero frequency filtering (ZFF) method. ZTL is an analysis technique which has capabilities to provide good temporal and spectral resolution for speech signals. It highlights the high SNR regions in the spectral domain. These advantages of the ZTL analysis technique served as a motivation for the development of the segmentation algorithm presented in this paper.

The paper is organized as follows: Sections 2 and 3 discusses the development of the ZTL technique and extraction of resonant frequency peaks. Section 4 presents the propoased segmentation method, the database used for evaluation, results and comparison with the existing techniques. Section 5 discusses the possible causes of errors in segment boundary detection as well as the further studies along these lines.

2. Zero Time Liftering : motivation and method

Zero time liftering (ZTL) [12] is a recently proposed method for the analysis of speech signals which provides high temporal resolution maintaining simultaneously a good spectral resolution. ZTL involves multiplying of the speech signal with a highly decaying impulse-like window, ensuring high resolution in time. The window function is given by

$$h[n] = \frac{1}{8\sin^4(\pi n/N)} \quad , \quad n = 0, 1, 2, ..., N-1 \quad (2)$$

This filter is used to multiply the signal $s[n]$ starting at a reference point n=0, and this imparts a polynomial-type growth/decay to DFT samples in the frequency domain. The hidden spectral features are highlighted by successive differencing of the numerator of the group delay (NGD) function, which is given by

$$g(\omega) = X_I(\omega)Y_R(\omega) - X_R(\omega)Y_I(\omega) \quad (3)$$

where $X(\omega) = X_R(\omega) + jX_I(\omega)$ is the DTFT of $x[n]$ and $Y(\omega) = Y_R(\omega) + jY_I(\omega)$ is the DTFT of $nx[n]$.The spectrum is represented by the Hilbert envelope of the NGD (HNGD) which has a good resolution around the formants[13]. ZTL analysis involves the windowing of speech signal using $h[n]$ with a shift of one sample to calculate the spectrum. The spectral characteristics for a signal can be obtained for segments starting at any instant of time, and hence the results can be interpreted as

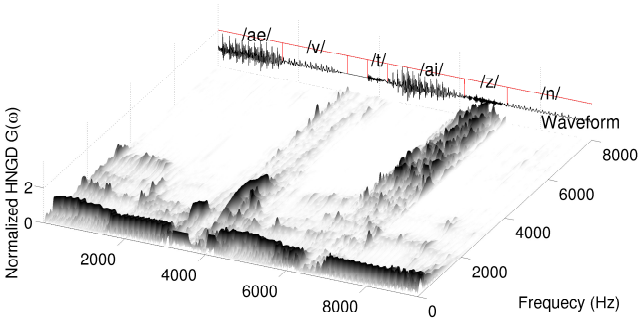


Figure 1: *Analysis of a speech segment using ZTL method and DRFs. The figure shows the speech waveform and the corresponding HNGD spectra after ZTL analysis.*

$instantaneous\ spectral\ features$. The energy profile of the ZTL spectra at every instant can be attributed mostly to the signal sample at that instant, and to a few other samples in the vicinity. Figure 1 shows a speech signal for the utterance *'advertising'* and the corresponding ZTL spectrum computed using a window length of 10ms at a sampling rate of $f_s = 16kHz$, i.e., for $N = 160$, and shifting this window by every sample. The capabilities of the HNGD spectra to provide a high temporal resolution and highlighting the spectral peaks for a given speech signal is well evident from the figure.

3. Obtaining DRF using ZTL and segmentation of speech

ZTL provides an insight to the production characteristics for different acoustic segments of speech signals. Analysis of speech using this method can help in understanding the difference in spectral behavior for segments corresponding to various acoustic events. Figure 1 shows the change in spectra corresponding to various acoustic events. The location of the spectral peaks and their corresponding strengths in the spectra led to the development of representation of speech signal using the most prominent peak. These spectral peaks correspond to high SNR regions which are less affected by environmental factors, and thus are robust in representing the speech signals.

The given speech signal is analyzed using ZTL and the frequency of the dominant peaks from the spectra with their amplitudes are obtained. When these frequencies are plotted along the signal, it was observed that there is clear distinction of various acoustic segments. These frequencies represent the dominant resonances of speech segment being analyzed, and can equivalently be associated with the dimensions of the prominent cavity in the vocal tract responsible for the production of that segment. These resonance peaks are thus called dominant resonant frequencies (DRFs), and are considered as representation of the production characteristics for a speech signal. Figure 2 shows speech signals corresponding to some acoustic events and their DRFs with their respective amplitudes. It can be observed from the figure that DRFs relate to the instantaneous production characteristics for a speech signal and thus provide evidence to identify distinct acoustic segments in the speech signal. The temporal resolution obtained by ZTL helps in plotting DRFs at every sampling instant.

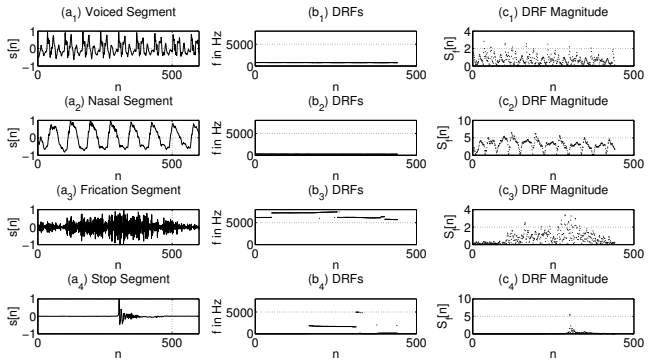


Figure 2: *Different acoustic segment of speech and corresponding DRFs along with their magnitude. Plots 'a' show speech segments for voiced, nasal, frication and stop respective and plots 'b' and 'c' show the corresponding DRFs and respective spectral magnitude.*

3.1. Distinct acoustic events and respective DRF behavior

The DRFs of the speech signal help demarcate boundaries for distinct acoustic events in the speech signal with good accuracy. The acoustic events correspond to the change in the shape of the vocal tract which produce different phonetic classes of sounds. Multi-dimensional representation of the source and system responses are used to build models to learn the changes during the production of the speech signal. These learning processes are supervised in nature, and thus require a large number of instances for each of such transitions. Employing DRFs on the other hand doesn't require any training but just the pre-acquired knowledge of production mechanism for different classes of sounds. **Voiced** segments, for instance, are produced with a cavity which is wide open without creating any constriction in the vocal tract. Whereas **obstruents** are produced by creating constrictions at various locations using different articulators, and this gives rise to different cavity shapes for these sounds. Such cavities, being smaller in length compared to the voiced sounds results an increase in the frequency of resonance. **Nasals** are other class of sounds which are produced by the coupling of the vocal and nasal tracts by lowering of the velum. This results in a cavity length longer as compared to vowel sounds and thus resonate at frequencies lower than vowels. **Plosives** are classes, where a closure in the vocal tract is followed by a burst which is impulsive in nature, and thus results in energy distribution in multiple frequency components with a short onset and offset. These and many more acoustic segment transitions are captured efficiently using DRF representation. Experiments were conducted to test the consistency of DRFs across different utterances and speakers. In the case of clean speech, DRFs proved to be quite robust and consistent in representing the production characteristics for the speech signals.

4. Database, method and segmentation results

The segmentation problem requires ZTL analysis to be performed on the given speech signals. An analysis with different window length, N $(= 2, 5, 10$ and $20ms)$ as given in Eq (2), was carried out for speech, and DRFs were obtained from the corresponding HNGD spectra. On examination of the DRF patterns, we choose $N = 10ms$ where the DRF representations appear smooth, and segmentation can be done easily.

The algorithm to perform the segmentation tries to identify the changes in the acoustic properties of the signal. To segment the speech signal, we first differentiate between obstruent and sonorant regions based on the characteristics of of the respective DRFs as explained in section 3.1. Further, a 3-$point$ median filtering is performed in the sonorant regions to smoothen the DRF curves. Changes in vocal tract cavity shape within sonorant regions can identified with a transition in location of DRFs. The transition parameter f_{tr} controls the number of segment boundaries being generated by the proposed algorithm. It is observed that a range of $f_{tr} = 30$ to $120Hz$ provides almost similar values of α. Multiple boundaries occuring within a window of $20ms$ are then merged to one to avoid ambiguity.

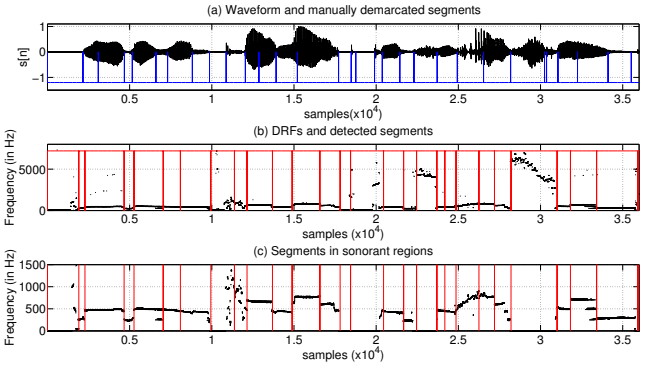


Figure 3: *Segmentation results based on DRF representation of speech signal. Plot 'a' shows a speech signal and the corresponding manually transcribed boundaries. Plot 'b' shows the corresponding DRFs and boundaries obtain with this representation. Plot 'c' shows the sonorant regions with corresponding boundaries with* 3-$point$ *median filtered DRF curves.*

The segmentation algorithm was evaluated on a subset chosen from the TIMIT database [14]. TIMIT is the most widely used corpus for phone segmentation task. It consists of microphone quality recordings of 630 American-English speakers (10 sentences per speaker), with sampling frequency $16kHz$ and resolution 16-bit. The chosen subset contains 182 sentences from TIMIT dataset, uttered by an equal number of male and female speakers. Figure 3 shows one such waveform with the manual transcription boundaries and the corresponding DRF representation of the signal with the boundaries generated by the proposed method. We can see that DRFs represents the changes in production characteristics in the given signal.

The performance of segmentation is reported by comparing the algorithmic segmentation with manual labels provided with the TIMIT database in terms of CDR(α). A tolerance window of $20ms$ is generally chosen to report the performances [8]. We computed the results for the segmentation using DRFs for tolerance levels of $5ms$, $10ms$, $15ms$ and $20ms$, which are shown in Table 1.

Another parameter comparing the performance of segmentation algorithms is the *over-segmentation rate* (*OSR*) β, for the algorithm. The percentage of over-segmentation is given as

$$\beta(in\%) = \left(\frac{N_{detected}}{N_{truth}} - 1\right) * 100, \qquad (4)$$

Table 1: *Performance of segmentation using DRFs on TIMIT database for different tolerance levels. The CDR is expressed in % and tolerance is expressed in ms.*

Tolerance	5ms	10ms	15ms	20ms
α	33.7	58.4	70.9	79.6

Table 2: *Comparison of segmentation results for a tolerance level of 20ms. The CDR(α) is expressed in % and tolerance is expressed in ms.*

Segmentation (supervised)	α
HMM+SVR	88.18
$(HMM+SVM)_1$	94.9
$(HMM+SVM)_2$	95.4
Segmentation (unsupervised)	α
MMC	67.9
AP_{seg}	77.2
DSF	77.2
DRF	79.6

where a negative β suggests the under-segmentation rate. Silence regions in speech signals have been reported as problematic for unsupervised speech segmentation algorithms [8] and therefore we ignored the segment boundaries within the silence regions for calculating β. For the rest of the segments boundaries, the proposed algorithm gives a β of around **11%**. As stated in the previous sections, the segment boundaries detected using DRFs correspond to acoustic changes, which sometimes may not correspond to any of the manually transcribed phoneme boundaries.

Table 2 shows the results obtained for the proposed method in comparison with results obtained by other methods over the same database. The comparison is made with respect to supervised as well as unsupervised segmentation techniques. For instance, HMM + SVR employs multiple base segmentation engines (BSEs) which are implemented with HMMs trained on different parameterization methods such as MFCC, LPCC, HFCC, PLP etc. and using *support vector regression* for boundary fusion. This method is explicit in nature, whereas both HMM + SVM methods are implicit, which is basically a segmentation method used for TTS systems. The $(HMM+SVM)_1$ and $(HMM+SVM)_2$ are similar models trained on the TIMIT database but tested on TIMIT and TTS datasets, respectively. They use trained phone models as HMMs and perform SVM (support vector machine) based refinement of local boundaries. Unsupervised methods like the maximum margin clustering, MMC, is a kernel based unsupervised form of SVMs which maximizes the separation margin between a set of unlabelled feature vectors. STM (Spectral transition measure) measures the magnitude of the spectral rate of change and DSF (delta spectral function) represents variation of band energy for a specific band for each frame. Phoneme transitions are usually reflected as peaks of such functions. The AP_{seg} method includes acoustic-phonetic features such as zero crossing rate ZCR, energy onset and offsets and formant energy ratio along with statistical learning methods to detect segment boundaries. When compared to the unsupervised methods, the proposed DRF based method gives better performance with a low β. There still lies some gap between performances of supervised and unsupervised segmentation methods which can be overcome by further refinements.

5. Error analysis and conclusions

The boundaries obtained by segmentation using the DRFs helps demarcating the acoustic events for a signal. These events boundaries signify the transition in vocal tract characteristics during the production of speech and ZTL analysis helps in marking these accurately. Yet Table 1 shows a low value of α at $5ms$ and $10ms$ tolerance levels. An error analysis was carried out over the segment boundaries obtained with DRFs and observation of the signal characteristics in the vicinity of these boundaries. This analysis suggests that the manual transcriptions might be wrongly placed in some cases. The event boundaries demarcation process with DRFs is based on the extraction of acoustic properties of the signal, and therefore are likely to be more accurate.

There are several advantages with the proposed method for representing the speech signals. This method is unsupervised, as it has no requirements in terms of the acquiring knowledge and learning from examples. The results obtained prove the capability of DRFs as consistent evidence to identify the acoustic boundaries of speech signals. Furthermore, this method is independent of language, gender and several other speaker and corpus based dependencies. The representation provided by DRFs for the segments in speech signal also helps in visualizing the boundaries manually, primarily for annotation purposes. Resolution of automatic methods depend on their frame sizes which is generally 10-20ms in size, whereas ZTL has a high resolution comparable to manual segmentation process.

Future work is planned to incorporate other speech production based features, which together with DRFs can help in improving the segmentation performance. It is also proposed to automatically label the acoustic segments into different categories.

6. References

[1] C. Cucchiarini, "Phonetic transcription: A methodological and empirical study," Ph.D. thesis, University of Nijmegen, Nijmegen, The Netherlands, 1993.

[2] J. P. van Hemert, "Automatic segmentation of speech," IEEE Transactions on Signal Processing, vol. 39, no. 4, pp. 1008–1012, 1991.

[3] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on Hidden Markov Models," Speech Communications, vol. 12, no. 4, pp. 357-370, 1993.

[4] D. T. Toledano, L. A. H. Gmez, and L. V. Grande, "Automatic Phonetic Segmentation," IEEE transactions on speech and audio processing, vol. 11, no. 6, pp. 617–625, 2003.

[5] Mporas Iosif, Ganchev Todor and Fakotakis Nikos, "Phonetic segmentation using multiple speech features," International Journal of Speech Technology, Springer Netherlands, vol. 11, no. 2, pp 73–85, 2008.

[6] Iosif Mporas, Todor Ganchev, Nikos Fakotakis, "Speech segmentation using regression fusion of boundary predictions", Computer Speech & Language, Volume 24, Issue 2, pp. 273–288, 2010.

[7] Juneja, A. and Espy-Wilson, C., "Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning.", in the proceedings of International Conference on Neural Information Processing, Singapore, November 18-22, 2002

[8] Odette Scharenborg, Vincent Wan, and Mirjam Ernestus, "Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries," Journal of the Acoustical Society of America, vol. 127, Issue 2, pp. 1084-1095, 2010.

[9] Pereiro Estevan, Y., Wan, V., and Scharenborg, O., "Finding maximum margin segments in speech," in Proceedings of ICASSP 2007, Honolulu, HI, 2007.

[10] D. T. Hoang, Hsiao-Chuan Wang, "Unsupervised phone segmentation method using delta spectral function," International Conference on Speech Database and Assessments (Oriental COCOSDA), pp.152–156.Hsinchu, Taiwan, 2011.

[11] Dusan, S., and Rabiner, L. "On the relation between maximum spectral transition positions and phone boundaries," in Proceedings of Interspeech 2006, Pittsburgh, USA, 2006.

[12] Dhananjaya N, "Signal processing for excitation-based analysis of acoustic events in speech", Ph.D. Thesis, IIT Madras, Chennai, India, October 2011.

[13] Anand Joseph M, Guruprasad S, and B. Yegnanarayana,(2006) "Extracting formants from short segments of speech using group delay functions," in Proceedings of Interspeech 2006, Pittsburgh, USA, 2006.

[14] Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., and Dahlgren, N. "Darpa timit acoustic-phonetic continuous speech corpus, Technical Report NISTIR 4930, National Institute of Standards and Technology, Gaithersburg, MD, 1993.

Syllable Nuclei Detection Using Perceptually Significant Features

A. Apoorv Reddy[1], *Nivedita Chennupati*[2], *B. Yegnanarayana*[3]

Speech and Vision Lab, IIIT Hyderabad, A.P, India

{[1]apoorv.reddy, [2]nivedita}@research.iiit.ac.in, [3]yegna@iiit.ac.in

Abstract

Speech can be segmented into syllables by identifying the syllable nuclei, which are points of high sonority. The excitation peaks in the linear prediction (LP) residual and the formant peaks can be interpreted as perceptually significant point features which contribute to the loudness of speech. In this paper, the use of these two point features is described for the use of detecting syllable nuclei. Each of these evidences contain information about different aspects of speech production, namely the glottal vibrations and the time varying vocal tract system. Thus it is possible that they contain complementary information about the syllable nuclei. Performance of the proposed syllable nuclei detection algorithm is evaluated for the TIMIT, Switchboard and the NTIMIT corpus. The proposed method performs comparably against two other state of the art syllable nuclei detection methods, and is shown to perform better for conversational speech. It is very fast and requires no training.

Index Terms: sonority, syllable nuclei detection, glottal closure instant, group delay function, LP residual

1. Introduction

At the perceptual level, the syllable nuclei are attributed to high energy sonorants or resonant sounds, which are relatively loud and carry a clear pitch. These attributes lead us to infer that the acoustic correlates of syllable nuclei are energy and periodicity properties [1]. An energy-based syllable detection method was proposed in [2], where energy peaks in the range 250 to 2500 Hz are shown to be well correlated with the syllable nuclei. A smoothed modified loudness contour is used to detect vowels for the purpose of estimating speaking rate in [3]. Speech rate estimation methods have mainly used the durations between the syllable nuclei as a method to find an estimate of the speaking rate [4, 5]. Syllable detection in [4] uses spectral correlation envelopes using selected sub-bands with temporal correlation and smoothing. Monte-carlo simulations are performed to find optimal settings for subband selection and the thresholds used for peak picking. A rhythm guided syllable detection algorithm is proposed in [5] where the rhythmic feature of the sequence of syllables in continuous speech is exploited. The parameters of an optimal sinusoid are calculated on the basis of peaks detected apriori in the energy envelope. A least squares fitting criterion is used to calculate the frequency and phase offset of the sinusoid based on the detected peaks. Then the next peak is detected in the energy envelope in a range around the next sinusoid peak. This range is dictated by the frequency of the sinusoid. The sinusoid is updated after calculation of each peak. A hierarchical hidden Markov model (HMM) based method is proposed in [6], which automatically syllabifies the input speech by generating *syl* and *garbage* tags for the input frame. A multilayer perceptron based automatic syllable boundary detection method is described in [7]. Here, the neural network tries to estimate the posterior probabilities of a phoneme being in syllable nuclear position in the context of neighbouring phonemes. Then possible errors are corrected automatically by parsing the decision output string which was obtained from the posterior probabilities of each phoneme. Wu et. al proposed the use of a multilayer perceptron based classifier to detect syllabic onsets which were subsequently shown to improve speech recognition [8]. In [9], a bidirectional long short-term memory neural network model is used to identify potential syllable nuclei in spontaneous and read speech. The neural network uses a 79 dimensional input vector including a 20 sub-band modulation spectrum, their first differences, 12 PLP coefficients, log energy and their first and second differences. The output was specified by Gaussian curves spanning the duration of the syllable nuclei, and setting the rest of the output as zero. The neural network was trained using the gradient descent algorithm.

The acoustic correlates used for detection of syllable nuclei are based on our limited understanding of the production features relating to the perception of the syllable, such as a high energy sonorant or a relatively loud sound carrying clear pitch. The concepts of perception of energy and pitch are due to some features derived from a finite duration segment of speech. For example, energy is generally computed over 20-30 ms, assuming stationarity of the vocal tract system during that interval. Likewise, pitch periodicity can be perceived only if the signal is processed over a few cycles of the glottal vibration. In fact the least periodic sounds like whispers, fricatives, affricates and stops do not correpsond to the syllable nuclei. Most voiceless sounds do not possess the characteristics of syllable nuclei.

In this paper, we examine a set of new acoustic correlates that can contribute to the perception of high energy sonorants and relatively loud sound for detecting syllable nuclei. The new acoustic correlates are based on the fact that in voiced speech, the primary source of excitation of the vocal tract system is by the impulse-like characteristics due to the sharp closure of the vocal folds in each glottal cycle. It is well known that sharper the closure, the louder is the speech sound [10]. This can happen without any relation to the periodicity or energy of the signal. Also, the vowels or sonorants are perceived louder due to the resonances of the obstruction free vocal tract. The sharper the resonances, i.e., lower the bandwidths, the louder is the corresponding sound. These two properties, the impulse-like excitation and low bandwidth resonances, can be interpreted as some kind of point properties, in the sense that impulse-like behaviour is confined to a very short ($<$ 1ms) region in the time domain and the sharp low bandwidth formants have the spectral energy concentration around the formant frequencies.

We hypothesize that the perception of high energy (loudness) and sonority could be due to the impulse-like excitation in time domain and sharp resonances in the frequency domain. Both these are point properties (as opposed to spread) in their respective domains. Subjective experiments on speech, synthe-

sized by suppressing and enhancing these two point features, confirm this hypothesis. Note that both these features are robust in the sense that they are local high SNR regions in their respective domains. We develop acoustic correlates that reflect these features, and show that they can help in identifying syllable nuclei.

In Section 2 we conduct subjective experiments to justify the hypothesis by modifying the excitation and resonance features in the signal. Section 3 gives an algorithm to detect the syllable nuclei. Section 4 gives the performance of the syllable nuclei detection method on TIMIT and Switchboard data. Robustness of the method is examined by evaluating its performance on the NTIMIT corpus. Section 5 gives a summary of the work reported in this paper.

2. Subjective experiments

In the speech signal the characteristics of the time varying vocal tract system can be represented approximately by the linear prediction coefficients (LPCs) derived for each frame (about 10-30 ms) of data using LP analysis. The LP residual represents some of the features of the time varying excitation. In particular, in voiced segments the impulse-like excitation is reflected as large energy of the residual signal around the glottal closure instants (GCIs). The impulse-like behaviour can be seen better in the Hilbert envelope (HE) of the LP residual. The sharpness of these peaks around the GCIs gives a perception of loudness [10]. The sharpness can be increased by multiplying the residual using a sequence of Gaussian-shaped pulses located around the GCIs. The modified LP residual is used to excite the time varying all-pole model represented by the LPCs for each frame.

On the other hand, to decrease the sharpness of the excitation peaks, each sample of the LP residual is divided by the square root of the corresponding sample in the Hilbert envelope of the LP residual.

The LPCs for each frame represent the shape of the vocal tract for that frame, and hence contain the information of the resonances or formants of the vocal tract system. The sharpness of the resonances in the LP spectrum may add to the perception of loudness caused by increasing sharpness of the peaks in the HE of the LP residual around the GCIs. To increase the sharpness of the peaks around the formant peaks, the speech signal (sampled at 8KHz) is passed through an all-pole filter represented by LPCs, which are derived from the LP residual signal of the speech signal obtained using a 1st order LP analysis. The first order LP analysis reduces the slope of the spectrum in the LP residual signal. A 7th order LP analysis of the 1st order LP residual gives LPCs which has peaks at the formant locations, however with a nearly flat overall slope. Hence passing the original speech signal through an all-pole filter represented by the LP7 coefficients emphasize the formants without changing the overall spectral slope.

On the other hand, the sharpness of the formant peaks can be decreased by passing the original signal through an inverse filter represented by the LP7 coefficients.

The original signal is thus modified in four ways, where the formant peaks are enhanced and de-emphasized, and the excitation peaks in the LP residual are sharpened and de-emphasized.

DR (de-emphasized residual) and EG refer (emphasized GCI) to the modified signals generated from the de-emphasized residual and the enhanced residual respectively. FS (formant suppressed)and FE (formant enhanced) represent the modified signal with suppressed formant peaks and emphasized formant peaks respectively.

	DR	*EG*	*FS*	*FE*
Sentence 1	-0.5	0.125	-0.625	0.75
Sentence 2	-0.5	0.125	-0.75	0.5
Sentence 3	-0.75	0.25	-0.75	0.75
AOS	**-0.583**	0.167	**-0.708**	**0.67**

Table 1: Average opinion scores for the modified signals with respect to the original signal.

Listeners were asked to listen to the original speech and the corresponding modified signals to mark the loudness level compared to the original signal. There were overall 3 sentences of 2-3 seconds duration each. Eight subjects were asked to give a score of +1 or -1 and 0 for the modified signals if they perceived the modified signals to be louder, muffled or the same as the original signal, respectively. The sentences were presented in the following manner: original, DR, EG, original, FS and FE. Table 1 gives the average opinion scores for the various modified signals on a scale of -1 to +1.

The the average opinion score (AOS) for the modifications for all three sentences calculated by averaging the opinion scores across all 8 listeners. The AOS indicates that there is a loss in perception of loudness if any one of the two, excitation source information or formant peaks are suppressed.

The subjective listening tests of the signal and modified speech signals indeed confirm that perception of loudness changes if any one of the two, excitation source information or the vocal tract system information is changed. The absence of the sharpness of the excitation peaks in the LP residual and the high bandwidth of the formant peaks in the modified signals give a perception of less loudness as compared to the original signal.

This can be seen in the spectrogram plots of the original signal as compared to the modified signals in Fig.1. The spectrograms in Fig.1 have been computed with a frame length of 20ms and a shift of 10ms. In Fig.1(b) and 1(c), the formant peaks have been suppressed and enhanced, respectively. The formant structure can not be easily observed in Fig.1(b), while it can be clearly seen in Fig.1(c). Fig. 1(d) and 1(e) correspond to the modified signals with a de-emphasized residual and enhanced residual, respectively. We can clearly observe from Fig. 1(d) and 1(e) that though the formant information is preserved, the formant magnitude is de-emphasized in 1(d) and emphasized in 1(e).

This observation gives us motivation to look at the excitation peaks and the formant peaks as acoustic correlates of loudness of speech, and thus to derive an envelope-based syllable nuclei detection method.

3. Envelope-based syllable nuclei detection

In this section we will describe the two evidences which are used to derive envelopes for syllable nuclei detection. They are based on the two point properties discussed before, namely, the excitation peaks in the LP residual and the formant peaks in the group delay spectrum.

3.1. Short time energy of the Hilbert envelope of LP residual (EHE)

A 14th order LP analysis is performed for each frame of 20ms with an overlap of 10ms at a sampling rate of F_s = 8kHz.

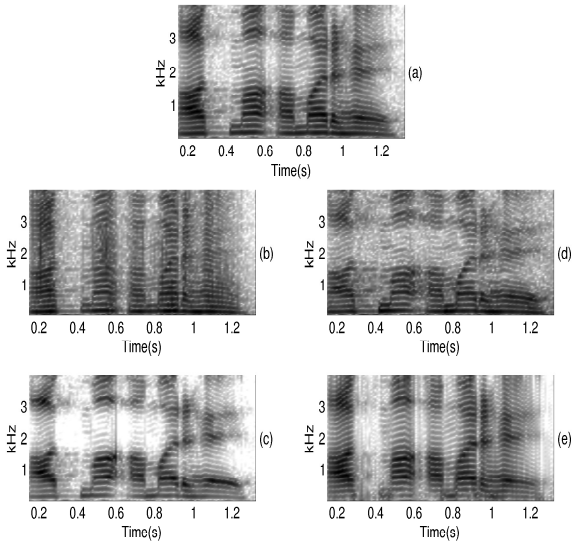


Figure 1: Spectrograms for the Hindi sentence *'mujhe nirantar ki ja:nka:ri acchi lagi'* corresponding to (a) Original speech signal (b) Formant suppressed (FS) (c) Formant Enhanced (FE) (d) De-emphasized LP residual (DR) (e) Emphasized LP residual (EG) .

The excitation peaks can be enhanced by computing the Hilbert envelope of the LP residual, as it serves to remove the phase information present in the excitation source [11]. The Hilbert envelope is the magnitude of the analytic signal $s_a(n)$ of the LP residual $e(n)$. The analytic signal $s_a(n)$ is,

$$s_a(n) = e(n) + je_h(n) \tag{1}$$

where $e_h(n)$ is the Hilbert transform of the LP residual $e(n)$, where $e_h(n)$ is calculated as follows:

$$e_h(n) = \begin{cases} \mathcal{F}^{-1}\{-j\mathcal{F}\{e(n)\}\}, \text{if } f \geq 0 \\ \mathcal{F}^{-1}\{j\mathcal{F}\{e(n)\}\}, \text{if } f < 0 \end{cases} \tag{2}$$

where f is the frequency and $\mathcal{F}$ denotes the Fourier transform.

The Hilbert envelope $h_e(n)$ of the LP residual $e(n)$ is computed as follows,

$$h_e(n) = \sqrt{e^2(n) + e_h^2(n)} \tag{3}$$

The GCIs of the speech signal are extracted using the zero frequency filtering (ZFF) method [12]. Energy of the HE of the LP residual is calculated around the GCIs with a window length of 1 ms and a shift of 1 sample. We take the local maxima of the energies calculated in this region as a measure of the loudness of the speech signal. We will call this energy profile as EHE.

To extract the syllable nuclei we need only observe gross level changes in the EHE. Thus to smear the local variations in the EHE we convolve it with a Hamming window of length 50ms. However a small peak corresponding to a syllable nucleus and lying in the neighbourhood of a relatively large peak will tend to get de-emphasized by this smoothing operation. Thus the square root of the EHE profile is taken before convolving it with the Hamming window.

3.2. Maximum formant magnitude envelope (MFME)

It is known that the group delay spectrum (GD) of a signal is proportional to the squared magnitude spectrum around the formant frequencies [13]. That is,

$$\tau_g(\omega) \propto |X(\omega)|^2 \tag{4}$$

where $X(\omega)$ is the Fourier transform of the signal $x(n)$.

The group delay function (GD) can be represented as a function of the real and imaginary parts of the signal spectrum in the following way [14],

$$\tau_g(\omega) = \frac{X_i(\omega)X_r'(\omega) - X_r(\omega)X_i'(\omega)}{X_r^2(\omega) + X_i^2(\omega)} \tag{5}$$

where,
$X(\omega) = X_r(\omega) + jX_i(\omega)$ and
$X'(\omega) = X_r'(\omega) + jX_i'(\omega)$ is the Fourier transform of the signal $nx(n)$.

The GD function is computed using the method described in [15] for each frame of 20 ms with a shift of 10 ms at a sampling rate of F_s = 8kHz. The formant peak with the highest magnitude in the GD spectrum will carry the most energy for that segment of speech. A contour is constructed by taking the formant peak in the GD spectrum with the maximum amplitude for each frame. The square root of this contour is taken as the MFME contour. The MFME contour is also smoothed by convolving it with a Hamming window of 50ms.

3.3. Combined evidence

Before combining the two evidences, we enhance each evidence. Each evidence is enhanced in the following manner. The first order difference (FOD) of the evidence is calculated. Spurious peaks in the individual evidences are eliminated using a simple slope counting method. The evidence is then amplitude normalized between two consecutive negative to positive zero crossings of the differenced evidence signal. The evidences EHE and MFME are then combined by taking their samplewise mean. This combined evidence is then normalized. A simple peak picking algorithm is used to find the local peaks. These peaks correspond to potential syllable nuclei.

Each peak in the individual evidences will have an amplitude of 1. So an amplitude threshold of 0.5 is used to remove spurious peaks in the combined evidence which fall below this threshold. A minimum spacing threshold of 75ms is used to remove a smaller peak if it lies in the neighbourhood of a larger peak. An adaptive thresholding technique described in [3] is used to further validate the detected peaks. For a peak, the combined evidence must fall below a threshold t, which is a fraction of the local maxima within a range D around the peak. We have taken $t = 0.8$ and $D = 75$ms. Peaks lying in unvoiced regions are removed by performing voice/unvoiced segmentation on the speech signal. Voiced/unvoiced segmentation of speech is done on the basis of the strength of excitation of the voiced epochs [16]. These spurious peaks may correspond to fricatives which have high energy.

Fig.2 illustrates the working of the syllable nuclei detection algorithm. The shaded regions correspond to the vowel regions in the sentence. The peaks in Fig.2(f) marked green are hits.

4. Evaluation

To evaluate the proposed method for syllable nuclei detection, the phonetically transcribed TIMIT and Switchboard corpora

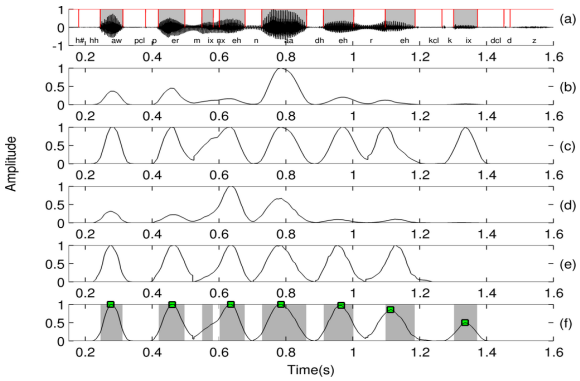


Figure 2: Syllable nuclei detection (a) Speech signal for the TIMIT sentence *'How permanent are their records?'*. (b) Maximum Formant Magnitude Envelope (MFME). (c) Enhanced MFME. (d) Energy of Hilbert Envelope of LP residual (EHE) (e) Enhanced EHE. (f) Combined Evidence. The shaded area correpsonds to the ground truth for syllable nuclei durations.

are used as the ground truths for the location of the vowel phones. If a detected syllable nuclei lies in the region of the vowel phone, it is marked as a hit. First, the performance of peak picking in the individual evidences on the TIMIT database is evaluated, and then compared with the performance of the combined evidence (COMB1).

In addition, the traditional energy contour is used to set a baseline for detecting syllable nuclei. The energy contour is smoothed by convolving it with a 50ms Hamming window and then normalized. The differenced energy contour is calculated. The energy contour is then normalized between the consecutive negative to positive zero crossings of the differenced energy contour.

For the purpose of validating the EHE evidence with the MFME evidence, another system is designed where the two evidences are combined by taking their product instead of their mean. We shall call this system as COMB2.

[%]	Energy	EHE	MFME	COMB1	COMB2
Recall	68.75	87.84	82.34	**92.67**	74.37
Precision	**98.53**	89.95	90.19	91.06	**96.02**
F-measure	80.99	88.88	86.09	**91.86**	83.82

Table 2: Comparison of syllable nuclei detection using the energy envelope, EHE envelope, MFME envelope and the combined envelopes COMB1 and COMB2.

The performance of the individual evidences and the combined evidence are tabled in Table 2. Recall is defined as the ratio of the number of hits to the number of syllable nuclei in the ground truth. Precision is defined as the ratio of the hits to the number of detected nuclei, while the F-measure is the harmonic mean of the recall and precision. The COMB1 system has a higher hit rate and a better F-measure than the COMB2 system, although its precision is lower. The energy contour guided method has the best precision, however its recall rate is very low. From Table 2, we can infer that in most cases, the individual evidences EHE and MFME reinforce each other, though they also provide complementary information to each other when one of the evidences is missing in a syllable nuclear position. This can also be seen in Fig. 2(c). The strength of

(a) **TIMIT**

[%]	RG	BLSTM	COMB1	COMB1 + Energy
Recall	86.59	92.22	**92.67**	**91.24**
Precision	98.86	95.82	**91.06**	**95.6**
F-measure	92.07	93.98	**91.86**	**93.37**

(b) **STP**

[%]	BLSTM	COMB1	COMB1+Energy
Recall	84.44	**87.98**	**85.7**
Precision	83.11	**83.31**	**85.47**
F-measure	83.74	**85.58**	**85.61**

Table 3: Comparison of rhythm guided syllable nuclei detection (RG), BLSTM syllabification method and proposed method for the TIMIT and Switchboard corpora.

the impulse like events in the error residual for the phone /ix/ in Fig. 2 is very low, which results in a small EHE evidence and is thus not considered for evidence normalization. For purposes of comparing with other syllable nuclei detection methods, we will consider the COMB1 system.

We compare our results against the state of the art speech rhythm guided syllable nuclei detection (RG) algorithm described in [5] and the bidirectional long-short-term memory neural network (BLSTM) syllabification method proposed in [9]. Table 3(a) and 3(b) compare the performance of the proposed method with RG and BLSTM for the phonetically transcribed TIMIT and Switchboard (STP) corpora respectively. We have not compared our method with RG for conversational speech as we couldn't find enough time to implement it on our own. The precision of the energy contour based syllable nuclei detection method has been exploited to subsequently reduce the false alarms. The individual evidences EHE and MFME are set to zero if the amplitude of the corresponding sample in the energy contour lies below a certain threshold. The proposed method performs comparably for the TIMIT database, but outperforms the BLSTM and RG methods for the Switchboard corpus. To test the robustness of the proposed syllable nuclei detection method, we have tested it on the NTIMIT database and obtained a recall rate of **91.46%**, precision of **79.11%** and an F-measure of **84.83%**.

5. Summary

The syllable nuclei positions are usually occupied by sonorants which are perceptually louder than other speech sounds. In this paper, we have explored two perceptually significant acoustic features which may be helpful for syllable nuclei detection. The short time energy of the HE of the LP residual and the formant peak information, both are features which are point properties in their respective domains, i.e., time and frequency. We have conducted subjective listening tests which validate that these point features are important for the perception of loudness in speech. These two features are used to generate a profile whose peaks may correpsond to potential syllable nuclei. We then evaluate the performance of our syllable nuclei detection method against RG [5] and BLSTM [9] and find that the results are on par with the current state of the art methods in case of the TIMIT corpus and significantly better for the Switchboard corpus. The advantage of our proposed method is that no training is required, and it is computationally fast.

6. References

[1] Z. Xie and P. Niyogi, "Robust acoustic-based syllable detection," in *Proc. ICSLP*, 2006.

[2] H. Pfitzinger, S. Burger, and S. Heid, "Syllable detection in read and spontaneous speech," in *Proc. ICSLP*, vol. 2, 1996, pp. 1261–1264.

[3] T. Pfau and G. Ruske, "Estimating the speaking rate by vowel detection," in *Proc. ICASSP*, vol. 2, 1998, pp. 945–948 vol.2.

[4] D. Wang and S. Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2190–2201, 2007.

[5] Y. Zhang and J. Glass, "Speech rhythm guided syllable nuclei detection," in *Proc. ICASSP*, 2009, pp. 3797–3800.

[6] P. Nel and J. du Preez, "Automatic syllabification using hierarchical hidden markov models," in *Proc. ICASSP*, vol. 1, 2003, pp. 768–771.

[7] J. Tian, "Data-driven approaches for automatic detection of syllable boundaries," in *Proc. ICSLP*. Citeseer, 2004.

[8] S.-L. Wu, M. L. Shire, S. Greenberg, and N. Morgan, "Integrating syllable boundary information into speech recognition," in *Proc. ICASSP*, vol. 2, 1997, pp. 987–990.

[9] C. Landsiedel, J. Edlund, F. Eyben, D. Neiberg, and B. Schuller, "Syllabification of conversational speech using bidirectional long-short-term memory neural networks," in *Proc. ICASSP*, 2011, pp. 5256–5259.

[10] G. Seshadri and B. Yegnanarayana, "Perceived loudness of speech based on the characteristics of glottal excitation source," *The Journal of the Acoustical Society of America*, vol. 126, pp. 2061–2071, 2009.

[11] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 4, pp. 309–319, 1979.

[12] B. Yegnanarayana and K. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 614–624, 2009.

[13] B. Yegnanarayana, "Formant extraction from linear-prediction phase spectra," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1638–1640, 1978.

[14] A. Oppenheim and R. Schafer, "Digital signal processing," *Prentice-Hall, Englewood Cliffs, NJ, 1975*.

[15] M. Anand Joseph, S. Guruprasad, and B. Yegnanarayana, "Extracting formants from short segments of speech using group delay functions," in *Proc. INTERSPEECH*, 2006, pp. 1009–1012.

[16] N. Dhananjaya and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," *IEEE Signal Processing Letters*, vol. 17, no. 3, pp. 273–276, 2010.

Acoustic analysis of trill sounds

N. Dhananjaya,[a] B. Yegnanarayana, and Peri Bhaskararao
International Institute of Information Technology, Hyderabad-500032, India

(Received 12 July 2011; revised 24 January 2012; accepted 26 January 2012)

In this paper, the acoustic–phonetic characteristics of *steady* apical trills—trill sounds produced by the periodic vibration of the apex of the tongue—are studied. Signal processing methods, namely, zero-frequency filtering and zero-time liftering of speech signals, are used to analyze the excitation source and the resonance characteristics of the vocal tract system, respectively. Although it is natural to expect the effect of trilling on the resonances of the vocal tract system, it is interesting to note that trilling influences the glottal source of excitation as well. The excitation characteristics derived using zero-frequency filtering of speech signals are glottal epochs, strength of impulses at the glottal epochs, and instantaneous fundamental frequency of the glottal vibration. Analysis based on zero-time liftering of speech signals is used to study the dynamic resonance characteristics of vocal tract system during the production of trill sounds. Qualitative analysis of trill sounds in different vowel contexts, and the acoustic cues that may help spotting trills in continuous speech are discussed.
 [http://dx.doi.org/10.1121/1.3688470]

PACS number(s): 43.72.Ar [SSN] Pages: 3141–3152

I. INTRODUCTION

Trills are a stricture type (Catford, 1977, p. 127), characterized primarily by 'the vibration of one speech organ against another, driven by aerodynamic conditions' (Ladefoged and Maddieson, 1996, p. 217). The most common trills involve the tip of the tongue vibrating against a contact point in the dental/alveolar region, and are called *apical trills* (McGowan, 1992; Ladefoged and Maddieson, 1996). Apical trills are the most common variety of trills among Indian languages. The objective of this paper is to derive the acoustic characteristics of apical trills from the speech signal. The effect of trilling on the glottal source of excitation and on the resonance characteristics of the vocal tract system is studied. The effect of different vowel context on the resonance characteristics of apical trills is also studied. The phonetic convention of the vowel context is indicated with a superscript to the base phoneme, such as [r^{a}] to denote a voiced apical trill [r] adjacent to the vowel [a]. In this paper, characteristics of the voiced apical trill [r] are studied in the context of three different vowels [a], [i], and [u].

The phonological aspects of trills, such as their occurrences in various world languages, and their relationship with other phonemes, have been reported by Maddieson (1984), Ladefoged and Maddieson (1996), and Ruhlen (1987). The production of an apical trill involves satisfying several articulatory, as well as aerodynamic constraints. The articulatory constraints concern the lingual and vocal tract configurations. The aerodynamic constraints concern the maintenance of the right amount of tension at the apex (tongue tip) and the requisite volume velocity of air flow through the stricture, which are essential for the initiation and sustenance of the apical vibration. The articulatory mechanics of tongue-tip vibration have been described by Catford (1977), Ladefoged and Maddieson (1996), Recasens (1991), and Spajic *et al*. (1996), and modeled by McGowan (1992). The aerodynamic characteristics and the phonological patterns of trills across languages are studied in detail by Solé (2002). Estimates of the transglottal (subglottal and supraglottal) pressure values with respect to the atmospheric pressure, and the pressure gradient across the lingual constriction, essential for initiating and sustaining voicing and trilling, respectively, have been obtained based on oropharyngeal pressure and oral air flow measurements. Solé (2002) has also studied some of the phonological patterns, such as the absence of nasal trills, preference for voiced trills, alternation and co-occurrence of trilling and frication, and trill devoicing, from an aerodynamic point of view.

Ladefoged and Maddieson (1996) have reported that acoustic trills in linguistic use usually consist of two to five periods, whereas apical trills typically consist of two to three periods of vibration (geminate occurrences may be longer). Based on spectrographic measurements made for Finnish and Russian apical trills, Ladefoged and Maddieson (1996) report a typical trill period of 50 ms (open and closed phases each of 25 ms duration), and hence a trilling rate of about 20 cycles in a second. Lindau (1985) reports a mean trilling rate of 25 Hz (18–33 Hz) measured over 25 speakers from seven different languages. An estimate of the trilling frequency of the tongue tip based on mechanical lumped element modeling of trill aerodynamics is given by McGowan (1992). The trilling rate of the tongue tip can be estimated using the formula $F_r = 1/(2\pi\sqrt{MC}) = 30$ Hz, where M is the mass of the tongue tip, estimated to be $\sim$1 g (by assuming an approximate surface area of the tongue tip involved in vibration to be 1 cm^2), and C is the mechanical compliance (inverse of stiffness) per unit area of the tongue tip (approximated to be 3×10^{-5} cm^3/dyne) (McGowan, 1992; Stevens, 1999). Several studies on phonemic trills in Spanish have been reported, such as categorization of the Spanish dialect continuum (Lipski, 1994), acoustic correlates to distinguish one

[a]Author to whom correspondence should be addressed. Electronic mail: dhanu@research.iiit.ac.in

phonemic trill from another (Colantoni, 2006), and acoustic characterization of trills (Henriksen and Willis, 2010). The acoustic correlates studied mostly concern the number of occlusions (or trill cycles) and the duration of the trill. Manual measurements of these parameters are made by observing the acoustic waveforms and the spectrograms. Based on these acoustic parameters, a detailed statistical analysis of trills in Spanish in terms of its sociolinguistic implications has been made by Diaz-Campos (2008) and Henriksen and Willis (2010). Studies made in deriving the acoustic–-phonetic characteristics of trills from speech data are limited by the standard spectrographic tools derived from short-time spectral analysis. The dynamic nature of the vocal tract system during production of trills is likely to have an effect on the excitation source due to coupling of the excitation source and the vocal tract system.

Currently available signal processing techniques may not be adequate to study the dynamic source and system characteristics of the trill sounds. In this paper, some recently proposed signal processing techniques, together with conventional methods, are examined for the study of dynamic characteristics of the excitation source and the vocal tract system resonances. As discussed in the later sections, the new analysis techniques provide an interpretation of the results in terms of production characteristics of the trill sounds.

Murty and Yegnanarayana (2008) have proposed an approach based on the zero-frequency filtering (ZFF) of speech signals for analysis of impulse-like characteristics in the excitation source. The ZFF-based approach gives a simple but effective method for detection of the instants of glottal closure (GCIs) or epochs in voiced sounds. The method also provides a measure of the strength of excitation at the epochs and the instantaneous fundamental frequency (Yegnanarayana and Murty, 2009; Murty *et al.*, 2009). The regions around the GCIs have high signal-to-noise ratio (SNR), and hence are useful as anchor points for analysis of the characteristics of the vocal tract system. Traditional short-time spectral analysis of speech involves processing the signal in blocks of 10–30 ms. Magnitude spectrum computed over block sizes less than 10 ms is not useful for analysis of the vocal tract system, due to issues caused by time-frequency resolution. Recently, a new technique called *zero-time liftering* (ZTL) of speech signals for analysis of resonance characteristics of the vocal tract system has been proposed (Dhananjaya, 2011). The ZTL technique provides high resolution of the spectral characteristics in temporal domain. Multiplication of the speech signal in time domain by an impulse-like window function provides the high temporal resolution. This is called the "liftering" operation analogous to the operation done in cepstrum analysis of speech (Bogert *et al.*, 1963). Good resolution of the spectral characteristics in frequency domain is achieved using the group delay analysis (Yegnanarayana, 1978; Yegnanarayana and Murthy, 1992; Joseph *et al.*, 2006), where group delay is defined as the negative derivative of the phase of the Fourier transform of the signal (Oppenheim and Schafer, 1975, p. 19).

The paper is organized as follows: Production characteristics of lingual trills, primarily the tongue-tip trills, are described in Sec. II. Section III describes the zero-frequency filtering-based analysis for extracting the features of the excitation source from the speech signal. Section IV describes the zero-time liftering technique for analyzing the spectral characteristics of trills. Analysis of trill sounds in terms of excitation source and vocal tract resonance characteristics is given in Sec. V. Characteristics of trills in different vowel contexts are examined. The acoustic features for spotting trills in continuous speech are discussed. Characteristics of the voiceless apical trill, as well as the voiced and voiceless labial trills are also examined in this section. Similarities and/or contrasts between apical and labial trills are discussed. A summary of the paper along with directions for further research is given in Sec. VI.

II. PRODUCTION CHARACTERISTICS OF APICAL TRILLS

In the production of an apical trill the apex is voluntarily positioned by the speaker to make a contact (Ladefoged and Maddieson, 1996, p. 218) with the corresponding upper articulator. Almost immediately, the pulmonic egressive airstream that is flowing into the oral cavity increases the pressure gradient across the stricture. Due to the fine interaction between "tongue-tension and volume-velocity of the air-flow" (Catford, 1977, p. 127), the apical stricture gets broken and the apex falls down to some extent releasing part of the positive pressure gradient in the oral cavity. Then due to the Bernoulli effect, the apex recoils to meet the upper articulator and forms the next event of stricture. Thus, the closure–opening cycle repeats itself a few times, and the total number of such cycles constitutes the complete trill. One such closure–opening cycle may be referred to as a "trill cycle," and the different phases of an apical trill are depicted in Fig. 1. The typical rate of trilling of the tongue is ~20–30 Hz, and can be measured from the acoustic waveform or the spectrogram (McGowan, 1992; Stevens, 1999; Lindau, 1985; Ladefoged *et al.*, 1977).

Trill sounds usually have at least two trill cycles for them to be discriminated from another category of sounds, namely taps ([ɾ]), which have one single movement of the tongue from any arbitrary position to the roof of the oral cavity and back, analogous to a trill cycle. Lindau (1985, p. 166) observed that "from an acoustic point of view, a trill can be regarded as a series of taps." On the other hand, as observed by Recasens (1991), and Recasens and Pallars (1999), an apical trill differs from an apical tap in the overall tongue body configuration. Based on electropalatographic

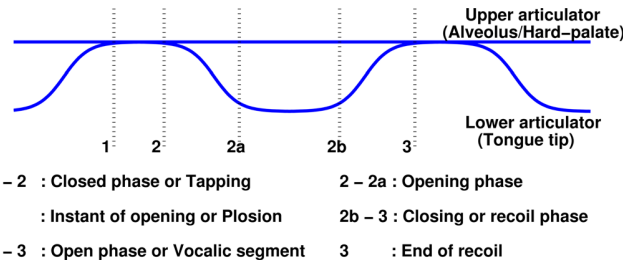


FIG. 1. (Color online) Illustration of different phases of the trill cycle for an apical trill. The durations and rate of change of articulator shown are only illustrative, and not actuals.

data, they observed that apical trills have a lower predorsum and a retracted postdorsum positions as compared to that of apical taps. Each of the constituent trill cycle is produced due to the Bernoulli effect, than due to voluntary movement of the apex of the tongue as in the case of a tap. But it should be noted here that the production of a trill still requires voluntarily maintaining the correct tongue body position, the right amount of tension (or stiffness) at the apex, and the requisite volume velocity of air flow across the stricture. In continuous speech two to three trill cycles are common (Lindau, 1985; Ladefoged and Maddieson, 1996; Henriksen and Willis, 2010), whereas in isolation they can be produced as a steady sustained sound with several (>3) trill cycles.

III. ZERO FREQUENCY FILTERING FOR ANALYSIS OF EXCITATION CHARACTERISTICS

Recently a zero-frequency filtering method was proposed for extracting the impulse-like characteristics of the excitation source from the speech signal, such as the GCIs, instantaneous fundamental frequency (F_0), and strength of excitation (Murty and Yegnanarayana, 2008; Yegnanarayana and Murty, 2009). The idea behind filtering the speech signal at zero frequency is that the effect of an impulse-like excitation source is equally felt throughout the spectrum, including around the zero frequency, whereas the vocal tract information is predominantly concentrated around the formant peaks. The method involves filtering the speech signal through a cascade of zero-frequency resonators. A zero-frequency resonator is an all-pole system with two poles at $z=+1$ in the z-plane, which is equivalent to a sequence of two cumulative sum operations in time-domain. This leads to a polynomial-type growth/decay of the output signal. The polynomial-type growth/decay can be removed by a trend removal operation, which involves subtracting the local mean from the signal at each time instant (Murty and Yegnanarayana, 2008; Yegnanarayana and Murty, 2009). The resulting signal is referred to as the *zero-frequency filtered signal*. The positive zero crossings (negative to positive) of the filtered signal correspond to the instants of glottal closure, also referred to as epochs. The slope of the filtered signal around the epochs gives a measure of the *strength of excitation*.

Figure 2(a) shows an example of a steady or prolonged trill sound uttered in isolation as a CV (consonant–vowel) unit [$r^a r^a a$] in the context of the vowel [a]. Note that in this paper, the repetition of the phone label [r^a] is used to denote the prolonged utterance of the phone, and the superscript denotes the vowel context. The output of cascade of two zero-frequency resonators, and the ZFF signal obtained after trend removal are shown in Figs. 2(c) and 2(d), respectively. The epoch locations given by the positive zero crossings of the ZFF signal [Fig. 2(d)] are shown in Fig. 2(a) by downward-pointing arrows. The strength of excitation (measured at the epoch locations as the slope of the ZFF signal) and the instantaneous fundamental frequency (measured as the reciprocal of the time interval between adjacent epochs) are shown in Figs. 2(e) and 2(f), respectively. The epoch locations occur at regular instants in most of the voiced regions (0.1–0.8 s), governed by strong impulse-like excitations imparted at the instants of glottal closure. The measured strengths of excitation at these epoch locations are also large for the voiced regions. In the silence regions (0–0.1 s and 0.82–0.92 s) and voiceless regions (not shown), the epoch locations occur at irregular instants due to lack of any regular impulse-like excitations, and the excitation strengths measured at these epochs are significantly lower compared to those in the voiced regions (Murty and Yegnanarayana, 2008). A simple threshold on the excitation strength helps to isolate the regions of voiced excitation. The voiced/nonvoiced decision based on the excitation strength is shown in Fig. 2(d). The ZFF-based method for extraction of epoch locations and their strengths has been shown to be robust against additive noise (Murty and Yegnanarayana, 2008; Dhananjaya and Yegnanarayana, 2010).

Trilling of the tongue tip affects the measured strength of the glottal excitation, as can be seen in Fig. 2(e). The strength of excitation varies within a trill cycle, and it is less during the closed phase as compared to the open phase. This may be due to the loading of the vocal folds by the closing of the oral cavity. It is also seen from Fig. 2(f) that the instantaneous fundamental frequency varies due to the trilling of the tongue tip. In contrast, the contours of the excitation strength and the instantaneous F_0 are relatively smooth within the vowel region (0.65–0.8 s in Fig. 2). A portion of the trill region of the waveform in Fig. 2(a) is shown expanded in Fig. 3 to show the details of the excitation characteristics of the trill. The fundamental frequency F_0 seems to reach a minimum value in the closed phase just before the release of apical contact, and increases gradually as the apical contact is forced open. At the same time, the F_0 movement toward the point of apical contact is not smooth, which probably hints toward a faster recoil of the apex than the opening as is observed in the case of vocal folds. More

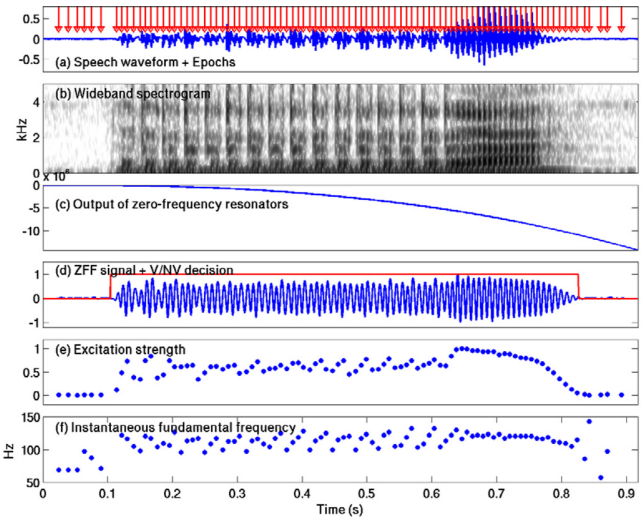


FIG. 2. (Color online) Zero frequency analysis of a steady or sustained apical trill produced as an isolated CV (consonant–vowel) [$r^a r^a a$]. (a) Speech waveform and the estimated epoch locations shown by downward arrows, (b) wideband (WB) spectrogram, (c) output of a cascade of two zero-frequency resonators, (d) ZFF signal after trend removal along with the V/NV decision, (e) excitation strength, and (f) instantaneous fundamental frequency (F_0).

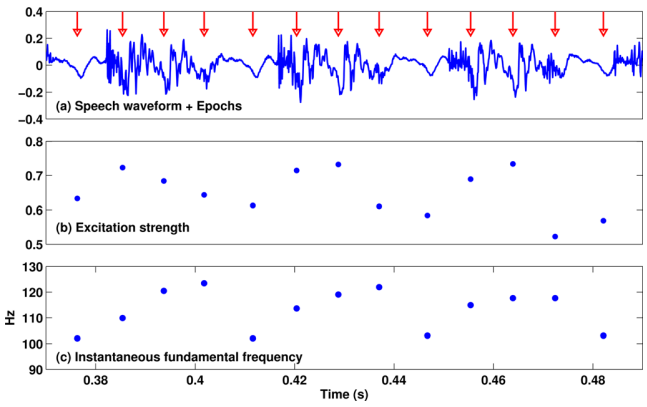


FIG. 3. (Color online) A portion of Fig. 2 expanded to illustrate the trill features. (a) Speech waveform along with estimated epoch locations, (b) excitation strength, and (c) instantaneous F_0.

evidence from other modalities such as eletroglottograph and/or magnetic resonance imaging (MRI) may be required to comment on the behavior of F_0 toward the point of apical contact. As per the aerodynamics of a pair of stretched membranes, analogous to the vocal folds, a minimum pressure gradient across the membrane is essential, depending on the mass and tension of the membrane, for the membrane to flutter or vibrate (Solé, 2002; Herman, 2007). As the subglottal pressure builds up behind the closed glottis, the pressure gradient between the subglottal and supraglottal air pressures increases, forcing the vocal folds to open with a burst of air rushing across the glottis. This results in a temporary reduction of the subglottal pressure, and hence reduction in the pressure gradient, allowing the vocal folds to recoil back to their initial stretched position due to the inherent myoelastic tension in the membrane. This cycle repeats itself. The pressure gradient has a direct relationship with the rate of vibration of the vocal folds, meaning, higher the gradient, higher the rate of vibration (van den Berg, 1957; Fant, 1960, p. 266). When there is a supraglottal oral constriction, as in the case of closed phase of the trills, the supraglottal oral pressure increases, reducing the pressure gradient across the glottis. This in turn may lead to the reduction in the rate of vibration of the vocal folds temporarily, which increases again gradually as the oral constriction is released, causing an increase in the pressure gradient across the glottis. Such a phenomenon is also reported by van den Berg (1957) and Fant (1960, p. 266) during the production of voiced occlusions. Assuming a constant lung effort and a constant tension in the vocal fold membranes, vibration of the vocal folds is directly influenced by the pressure gradient across the glottis, which in turn is influenced by the trilling of the tongue tip. The pressure gradient decreases during the closed phase, thus reducing the rate of vibration of the vocal folds. The pressure gradient increases during the open phase, resulting in an increase in the rate of vibration of the vocal folds. Reduction in the excitation strength during the closed phase can be explained by the reduced air flow due to a reduction in the transglottal pressure gradient, as observed by Westbury (1983) in the case of voiced occlusions. It can be seen from Figs. 2(e) and 2(f) that the fluctuating pattern in the excitation strength and in the instantaneous F_0 repeats itself over a few glottal cycles.

IV. ZERO-TIME LIFTERING FOR ANALYSIS OF DYNAMIC FEATURES OF VOCAL TRACT SYSTEM

Recently, a new method for analysis, called *zero-time liftering* of speech signals was proposed (Dhananjaya, 2011). The method provides high temporal resolution, simultaneously maintaining a good spectral resolution. Liftering of speech signal in the time domain with a heavily decaying impulse-like window provides high temporal resolution, whereas the group delay analysis provides good resolution of the spectral characteristics. The use of a heavily decaying liftering function smoothes the spectrum severely resulting in a polynomial-type growth/decay, analogous to that in the zero-frequency filtering (Murty and Yegnanarayana, 2008). The masked or hidden spectral features can be highlighted by successive differencing of the numerator of the group-delay function. Phase inconsistencies of some weak higher formants are handled by computing the Hilbert envelope of the differenced numerator of the group delay function (Joseph *et al.*, 2006). The resulting spectrum is referred to as *HNGD function*.

Figure 4 shows the HNGD plots computed at every sampled time instant for a prolonged utterance of the trill [$r^a r^a$] in the context of following vowel [a]. The speech waveform and the instants of glottal closure (downward

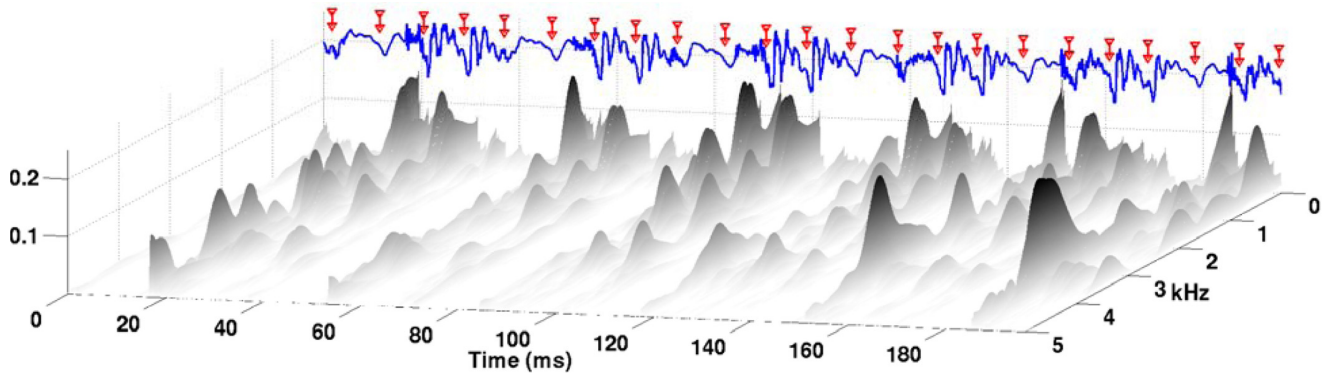


FIG. 4. (Color online) Waveform of a steady apical trill [$r^a r^a$] and the corresponding HNGD plots computed for every sample shift. The epoch locations are marked by downward arrows above the waveform.

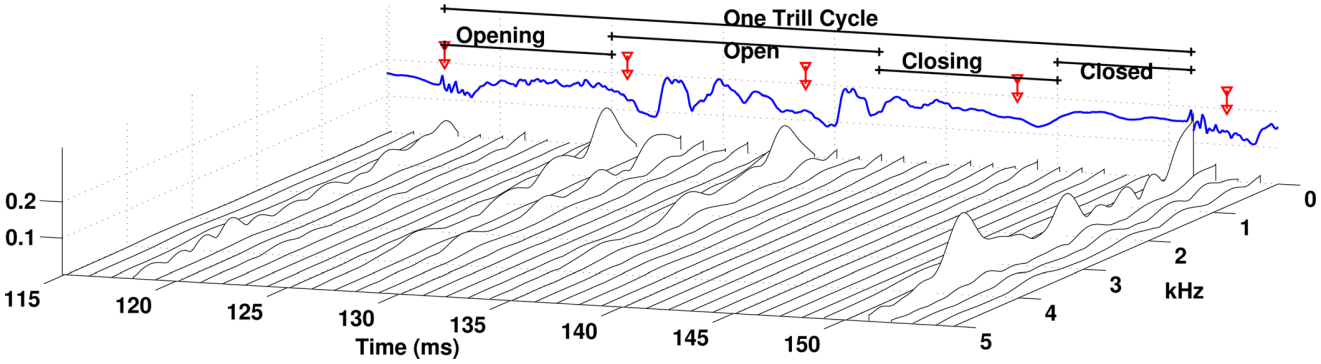


FIG. 5. (Color online) Waveform for one trill cycle of an apical trill [r^{a}r^a], and the corresponding HNGD plots computed for every 1 ms shift.

arrows) are also shown along with the HNGD plots for reference. The HNGD plots are computed over segments of length 4 ms ($M = 40$ at the sampling rate $F_s = 10\,000$), and using a discrete Fourier transform length of $N = 2048$. The segment of speech shown in Fig. 4 has approximately five trill cycles over 200 ms of duration, which is equivalent to a trilling frequency of ~25 Hz. The SNR of the speech signal varies continuously with time, and it can be seen that the HNGD plots around the high SNR regions (i.e., around the instants of glottal closure) are large compared to the HNGD plots in other portions of the signal. The time-varying nature of the spectral features can be seen better in Fig. 5, which shows one trill cycle of the speech signal and its HNGD plots computed at intervals of 1 ms. The GCIs are marked as downward pointing arrows. It can be seen from the waveform that at the beginning of the opening phase (~118 and ~150 ms) there is a burst along with a bit of frication due to the sudden opening of the tongue tip. The effect of the burst and the frication can be seen in the spectrum as a large peak around 3 kHz. The burst is more prominent around the time instant 150 ms compared to that around 118 ms, depending on the synchronization between the instant of opening of the tapping and the instants at which the spectrum is computed. This shows that one may fail to capture the instantaneous changes in the spectral characteristics, if the signal is sampled only at the epochs or even at a finer sampling rate of every 1 ms. Also, the trill sound seems to have characteristics of a voice bar [predominant low frequency band around the fundamental frequency without any significant formant structure as in the case of voiced occlusions (Dhananjaya *et al.*, 2008; Clark *et al.*, 2007, p. 278)] during the closed phase, which is partly apparent from the signal (in the region 145–150 ms), but cannot be seen in the HNGD plots of Fig. 4. The large dynamic range between the HNGD plots in the open phase and the closed phase within a trill cycle, and between the closed and open phases of the glottal cycle (region around the GCI), makes it difficult to observe the dynamic spectral characteristics of all regions of a trill sound simultaneously. One way of observing the instantaneous dynamic nature of the trill sounds is by normalizing the HNGD plots computed at every time instant. Figures 6(a) and 6(b) show the HNGD plots for one trill cycle with and without normalization. In Fig. 6(b) the HNGD plots are normalized by dividing each plot by its maximum value, so that all the HNGD plots are now in the range of 0–1. The instantaneous or time-varying spectral characteristics of the trill sounds, such as the large spectral peaks around 3.5 kHz (~118 and ~150 ms) due to bursts, and the voice-bar-like characteristics ~145 ms, can be observed better in Fig. 6(b) compared to Fig. 6(a).

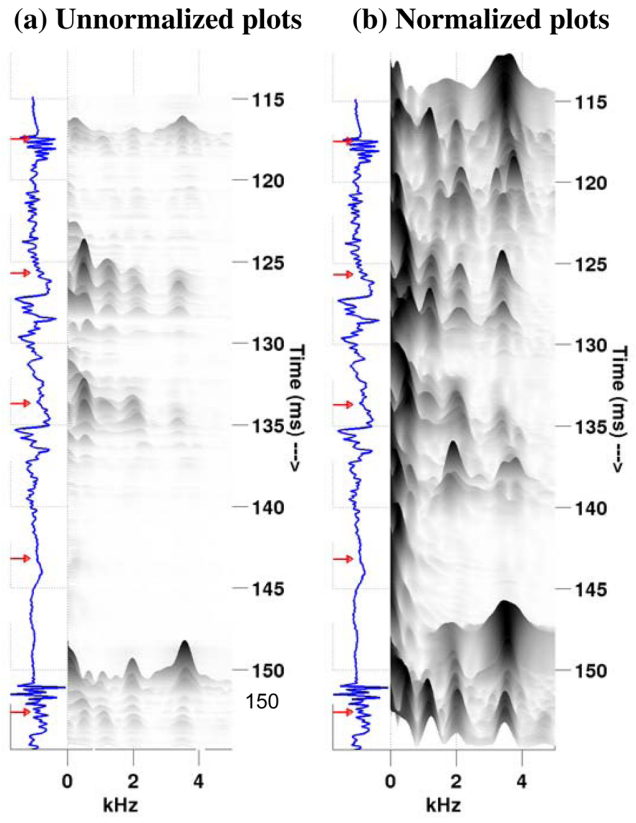


FIG. 6. (Color online) Waveform for one trill cycle of an apical trill [r^{a}r^a] and the corresponding HNGD plots. (a) Unnormalized HNGD plots and (b) normalized HNGD plots.

V. ANALYSIS OF TRILLS IN CONTINUOUS SPEECH

Like other continuants, trills are influenced by the adjacent vowel(s). In this section we examine the characteristics of a trill sound in the context of vowels [a], [i], and [u], which form the vertices of the vowel triangle in the F_1–F_2 formant space. Trills can also undergo transitions from one

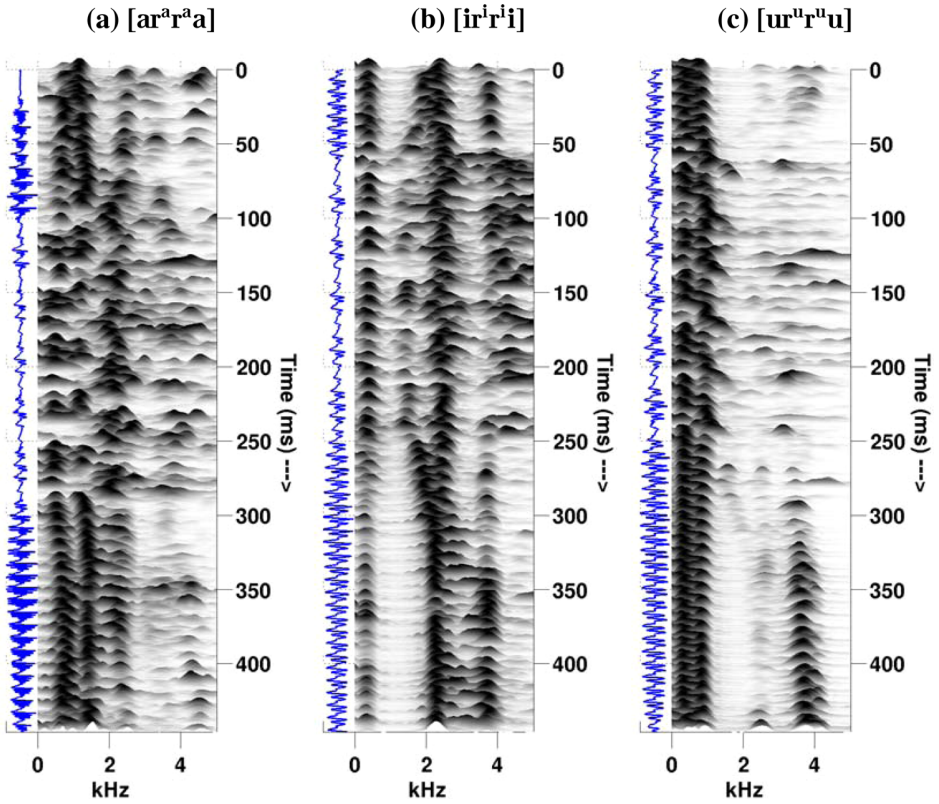


FIG. 7. (Color online) Waveform and HNGD plots of the apical trill [r] in the context of three different vowels [a], [i], and [u]. (a) [ar^{a}r^{a}a], (b) [ir^{i}r^{i}i], and (c) [ur^{u}r^{u}u].

vowel context to another, when they occur in between two vowels. Other categories of trills, namely, voiceless apical trills and bilabial trills, are also examined in comparison with the voiced apical trills.

A. Effect of vowel context on trills

A comparative study between the tap [ɾ] and the apical trill [r] by Recasens and Pallarés (1999) shows that trills are less affected by adjacent vowels as compared to taps, which is mainly due to the more constrained lingual position required for the production of an apical trill as against a tap. Recasens and Pallarés (1999) studied the coarticulation effects of apical trills with adjacent vowels using electropalatographic data, as well as formant frequency data, to show that a trill cannot be considered as a geminate correlate of a tap. Nevertheless, apical trills can be produced with varying vocal tract configurations irrespective of the small degree of freedom for variability. In this section, apical trills produced by a trained phonetician (male) in the context of three vowels [a], [i], and [u] are used to study the effect of vowel context on the trill. The three vowels provide three distinct vocal tract configurations.

Figure 7 shows the HNGD plots for trills uttered in three different vowel contexts, [ar^{a}r^{a}a], [ir^{i}r^{i}i], and [ur^{u}r^{u}u]. The trill sounds have been uttered as isolated VCVs (vowel–consonant–vowel), where the consonant C is the trill [r] and the vowel V is one of [a], [i], and [u]. The HNGD plots clearly show that the spectral peaks in the region of trill sounds are different in each of the three vowel contexts. This shows that the production of trill sounds need not have a unique vocal tract shape, although it has a highly constrained lingual configuration. Another observation that can be made from the HNGD plots is that the resonances of the trill sounds are more aligned with that of the vowels [a] and [u], as compared to that of [i]. This may be because the front-high tongue dorsum position for [i] needs to be retracted back considerably for the production of trill, which can be seen in terms of a decrease in second formant from [i] to [r]. The amount of reconfiguration required by the tongue body in the transition from a vowel to trill can be observed in terms of the changes in the vocal tract resonances. Figure 8 shows the locations of the context-dependent trills in the

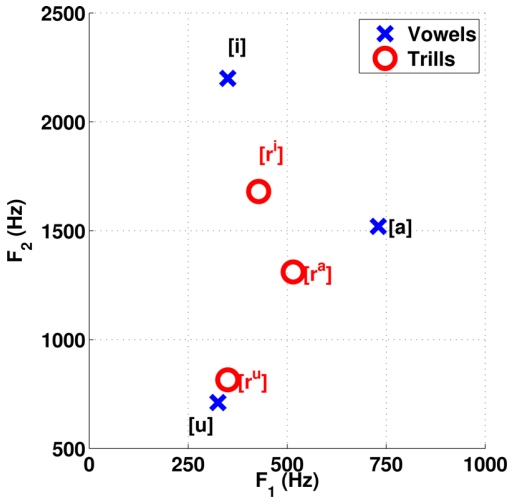


FIG. 8. (Color online) Vowel triangle formed by the vowels [a], [i], and [u], and the relative positions of the corresponding trills [r^{a}], [r^{i}], and [r^{u}], respectively. The F_1 and F_2 values shown here are the mean values, obtained by averaging the F_1 and F_2 values estimated at every 1 ms shift from a single utterance of the VCVs [ar^{a}r^{a}a], [ir^{i}r^{i}i], and [ur^{u}r^{u}u] by a trained male phonetician.

F_1–F_2 (first two formants) space, relative to the vowel triangle formed by the three vowels.

It can be seen that the change in the first two formants is minimal for a trill in the context of [u], whereas it is maximum for [i]. The largest F_1 movement (around ~250 Hz) between a vowel and a trill is observed in the context of [a] as the vocal tract changes from an open to a more closed position. Similarly, the largest F_2 movement (~500 Hz) is observed in the context of [i] as the frontal tongue dorsum position for [i] is retracted for the production of the trill. Comparison of trills produced in the three different vowel contexts (marked as circles in Fig. 8) shows that F_1 has a narrow spread of ~150 Hz (from approximately 350 to 500 Hz), whereas F_2 has a broader spread of ~900 Hz (from approximately 800 to 1700 Hz). The small variation in F_1 shows that the length of the vocal tract does not vary much, and is the longest (lowest F_1) in the context of vowel [u]. The large variations in F_2 may be attributed to the flexibility in the tongue dorsum position for the production of trills, with a highly retracted or back position in the context of [u] producing the lowest value of F_2. This observation based on the acoustic data may probably be verified by an analysis of data from other modalities, such as electropalatography, x-ray, and/or MRI. Although apical trills tend to take on the spectral characteristics of the adjacent vowel to a certain extent, they also tend to move toward a common space, due to the inherent articulatory constraints in their production.

Figures 7 and 8 show the characteristics of trills produced with a fixed vocal tract configuration (one of the three vowels [a], [i], and [u]) on either side. Trills can also be produced with a continuously changing vocal tract when the vowels on either side of the trill are not the same. Figure 9 shows the spectral characteristics of trills uttered as isolated V_1CV_2 units, where C denotes the apical trill [r], and V_1 and V_2 ($V_1 \neq V_2$) are one of the three vowels [a], [i], and [u]. The continuous transition of the spectral peaks from one vowel context to another can be clearly seen. In the case of trills transiting between vowels [a] and [i] {[ar^{a}r^{i}i] and [ir^{i}r^{a}a] as in Figs. 9(a) and 9(b)}, the key feature is the exaggerated movement of the second formant at the boundary between [r^i] and [i], whereas the transition of formants between [r^a] and [a] is more gradual. This is mainly due to significant reconfiguration required in the tongue body position between [i] and [r], as seen from Figs. 7 and 8. In the case of trills transiting between [u] and [a] {[ar^{a}r^{u}u] and [ur^{u}r^{a}a] as in Figs. 9(c) and 9(d)}, the key feature is the absence of any significant movement in the formants, as opposed to the case between [a] and [i], or [i] and [u]. Again trills transiting between [i] and [u] {[ir^{i}r^{u}u] and [ur^{u}r^{i}i] as in Figs. 9(e) and 9(f)} indicate a clear but gradual movement of the tongue body, as can be construed from the gradual movement of F_2.

B. Features for spotting steady trills in continuous speech

Acoustic cues for spotting steady trills in continuous speech are explored in this section. Unlike isolated utterances of trills, the majority of the trills in continuous or spontaneous speech tend to have fewer (less than three) trill cycles. They may have either one or two trill cycles, or the trilling may be totally absent at times, with the resulting sound being an approximant. Analysis of trills in the previous section shows the dynamic nature of the spectral characteristics of the trills, which vary with the vowel context. Hence, representation of the spectral characteristics of trills for spotting in continuous speech is a difficult issue. For spotting steady trills in continuous speech, an approach based on acoustic–phonetic knowledge using the excitation source characteristics seems to be more appropriate than a statistical approach using the spectral features. The excitation features that are useful in spotting the trills are the excitation strength and the instantaneous fundamental frequency. It can be recalled, Fig. 2 and Sec. III, that the excitation strength and the instantaneous F_0 vary in the trill region, whereas these parameters are almost steady and smooth for other voiced sounds. Figure 10(a) shows the waveform of a short utterance in Telugu, an Indian language, containing an apical trill. It can be seen that there are only two trill cycles in the utterance around the time instance 0.4 s (from approximately 0.35 to 0.45 s). Figures 10(d) and 10(e) show that the fluctuations in the excitation strength and the instantaneous F_0 can be observed for the trill sounds in continuous speech, which may be useful for distinguishing these regions from other voiced regions.

The speech waveform for a trill sound reflects two kinds of periodicity—a longer periodicity originating from the trill cycles and a shorter periodicity reflecting the glottal cycles. The presence of these two periodicities can be used as an additional cue for spotting trills in continuous speech. Normalized cross-correlation (NCC) can be used to measure the periodicity in a given signal. The NCC values for a given sequence $x[n]$, starting at an arbitrary time instant n, are computed as

$$\rho[k] = \frac{\sum_{n=0}^{M-1} x[n]x[n+k]}{\sqrt{\left(\sum_{n=0}^{M-1} x^2[n]\right)\left(\sum_{n=0}^{M-1} x^2[n+k]\right)}}, \quad (1)$$

where M is the window size over which NCC is computed and k is the shift or time-lag. The NCC values computed for a segment of trill are plotted in Fig. 11(b). The window size M used is 30 ms (240 samples at a sampling rate of 8 kHz). It can be seen that the NCC plot has multiple peaks. The highest peak (marked as TC—trill cycle), excluding the peak at 0 time-lag, corresponds to the periodicity due to apical vibration, and the first peak (marked as GC—glottal cycle) corresponds to the periodicity due to glottal vibration. Note that for a typical voiced segment, the highest peak in the NCC plot [Fig. 11(d)] is the peak due to glottal periodicity. The highest peak value $\rho_{\max}$ in the NCC plot is given by

$$\rho_{\max} = \max_k \{\rho[k]\}, \quad k = [N_1 : N_2], \quad (2)$$

where N_1 and N_2 are the lower and upper limits (in number of samples) for finding the highest peak, and may correspond typically to 2 and 50 ms, respectively. The time-lag $N_{\max}$ (in number of samples) of the maximum NCC value $\rho_{\max}$, and the corresponding frequency of periodicity $F_{\max}$ in Hz are given by

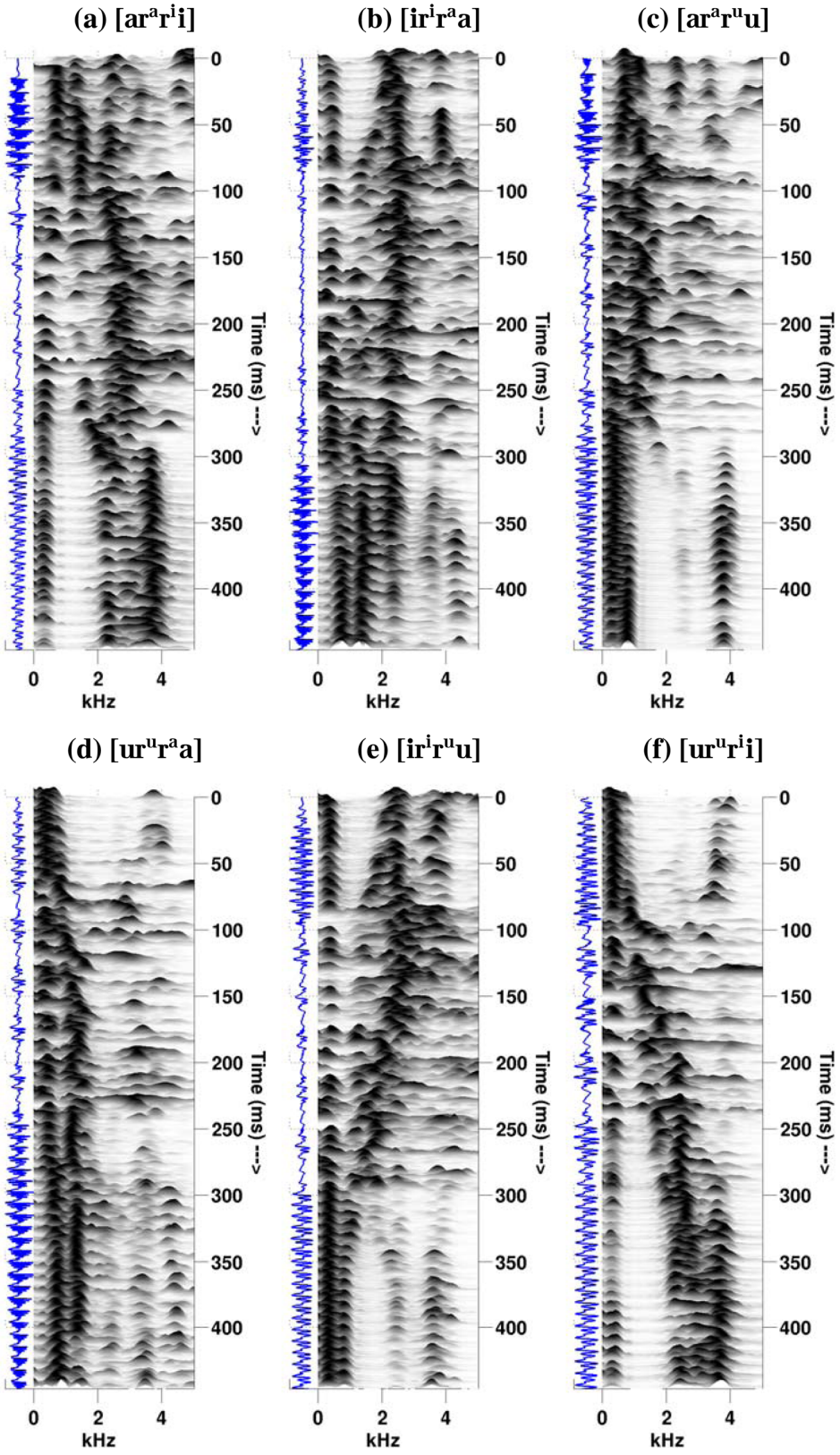


FIG. 9. (Color online) Waveform and HNGD plots of transitional trills changing from one vowel context to another. (a) [ar^{a}r^{i}i], (b) [ir^{i}r^{a}a], (c) [ar^{a}r^{u}u], (d) [ur^{u}r^{a}a], (e) [ir^{i}r^{u}u], and (f) [ur^{u}r^{i}i].

$$N_{\max} = \arg\max_{k}\{\rho[k]\}, \quad k = [N_1 : N_2] \tag{3}$$

and

$$F_{\max} = \frac{F_s}{N_{\max}}, \tag{4}$$

where F_s is the sampling frequency.

Figure 10(f) shows the $F_{\max}$ values computed for segments of speech starting at time instants n corresponding to the GCIs, using Eqs. (1), (3), and (4). It can be seen that the $F_{\max}$ values are low ($\sim$30 Hz) in the region of trill, whereas they are close to the instantaneous F_0 values in Fig. 10(e)

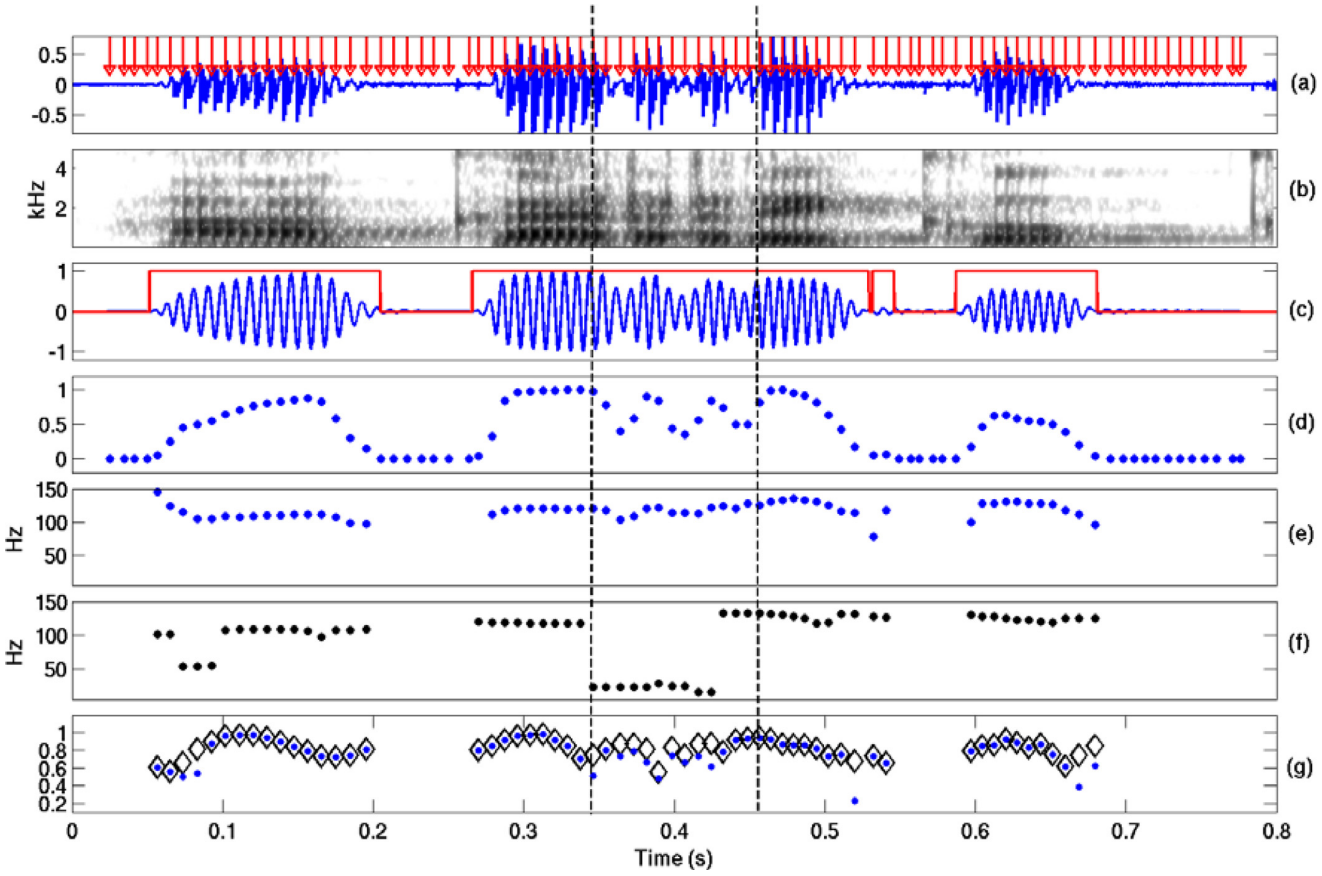


FIG. 10. (Color online) Analysis of trills in continuous speech for spotting. (a) Speech waveform of an utterance containing a trill [r] (0.35 to 0.45 s). The instants of glottal epochs are marked by downward arrows. (b) Wideband spectrogram, (c) ZFF signal with V/NV region marked, (d) excitation strength, (e) instantaneous F_0, (f) F_{max}—periodicity (in Hz) corresponding to the maximum NCC value ρ_{max}, and (g) NCC values ρ_{max} (diamonds) and ρ_0 (dots), corresponding to the highest peak and the peak around glottal pitch period, respectively.

obtained using the ZFF signal. Pitch halving or doubling is a common problem when correlation measures are used to estimate the periodicity in speech signals, especially in steady voiced regions. It can be seen in Fig. 10(f) that, at around the time instance 0.1 s, the F_{max} value is half that of the instantaneous F_0. Such spurious estimates can be identified using the knowledge that the contours of the excitation strength and the instantaneous F_0 are smoother in the steady voiced regions.

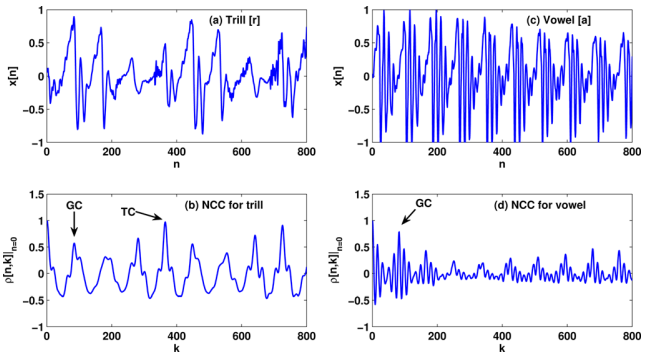


FIG. 11. (Color online) Normalized cross-correlation functions for a typical trill ([r]) and a voiced segment (vowel [a]). GC denotes correlation due to glottal cycle (maximum NCC value within the range of 2–20 ms), and TC denotes correlation due to trill cycle (maximum NCC value within the range of 2–50 ms).

Figure 10(g) shows the maximum NCC values ρ_{max} (marked as diamonds). The peak NCC values ρ_0 corresponding to the glottal periodicity are marked as dots, and they are obtained by locating the highest peak in $\rho[k]$ in the range of 2–20 ms. Note that the upper limit of the range is less than the period of the trill cycle. It can be noticed that ρ_0 and ρ_{max} values are exactly the same in most voiced regions, as the peak of NCC values around the glottal period also happens to be the highest peak. In the region of trill, the ρ_0 values are smaller than the ρ_{max} values, which correspond to the periodicities due to glottal and trill cycles, respectively.

C. Voiceless trills

Trills produced with the absence of glottal excitation (or voicing) are referred to as voiceless trills. Analogous to their voiced counterparts, voiceless trills can be produced in different vowel contexts. The spectral characteristics of the voiceless trill $[\underset{\circ}{\mathrm{r}}^{\mathrm{a}}]$ in the context of [a] are shown in Fig. 12(b). It should be noted that the first formant (typically around 600 Hz) for the vowels [a] (between 30 and 100 ms and 370 and 450 ms) is not clearly visible. In the first [a] (between 30 and 100 ms), the first formant is faintly visible but appears to merge with the second formant at ~1500 Hz. In the second [a] (between 370 and 450 ms), the first formant is completely invisible, whereas the second and third

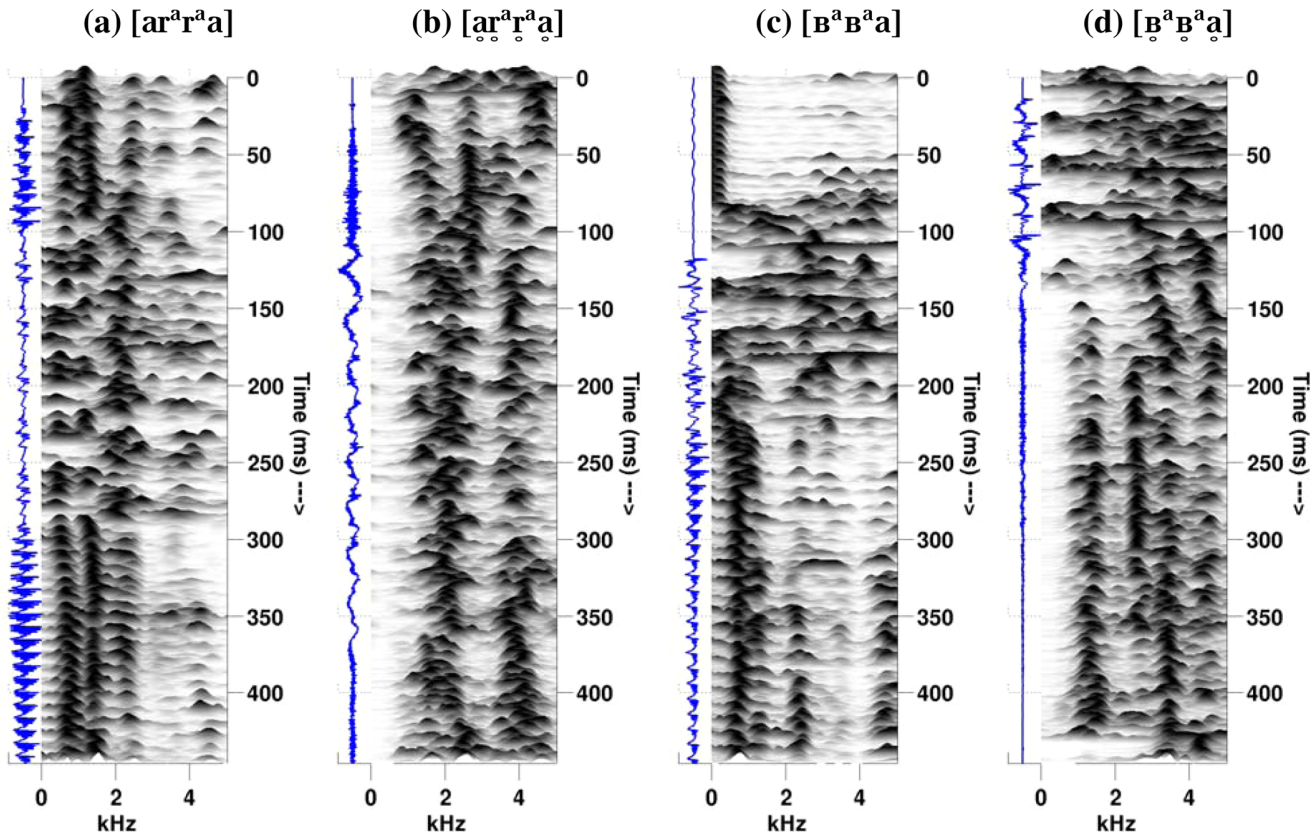


FIG. 12. (Color online) Waveform and HNGD plots for VCV units containing (a) voiced apical trill ([arᵃrᵃa]), (b) voiceless apical trill ([a̰r̰ᵃr̰ᵃa̰]), (c) voiced labial trill ([ʙᵃʙᵃa]), and (d) voiceless labial trill ([ʙ̰ᵃʙ̰ᵃa̰]), all uttered in the context of vowel [a].

formants seem to be merging. This may be due to the presence of a strong turbulent noise-like excitation, which typically has less energy in the low frequency region. Similar characteristics are observed in the case of other vowel contexts as well. Figure 13 shows the acoustic cues discussed in the previous sections for spotting voiced trills in continuous speech. The figure shows a voiceless trill [r̰ᵃ] followed by two bilabial trills, which will be discussed in the next section. It can be seen that the energy of the zero-frequency filtered signal [Fig. 13(c)] is significantly high in the region of voiceless trills, compared to the energy of the voiceless vowel [a̰] on either side. This is primarily because of the presence of impulse-like excitations generated by the trilling of the tongue tip. Hence, it is seen that the measured excitation strengths are large [Fig. 13(d)], and are detected as voiced region [Fig. 13(c)]. The epoch locations estimated within the voiceless trill are random, due to the noise-like characteristic of the excitation, and are irregularly spaced compared to a typical voiced region (vowel [a] between 1 to 1.2 s in Fig. 13). A similar trend can be observed for the instantaneous F_0 in Fig. 13(e). The signal periodicity F_{max} measured using the normalized cross-correlation is low (around 30 Hz) in the trill region as compared to other voiced regions (around 100 Hz), as can be seen in Fig. 13(f). It is to be noted that there is doubling of the estimated periodicity (or halving of F_{max}) at some instants toward the beginning of the trill. This can happen due to the presence of three or more steady trill cycles. This by itself may not be an issue for spotting voiceless trills, but needs to be addressed if an accurate estimate of the trill frequency is required.

D. Labial trills

Production mechanism of bilabial trills (voiced—[ʙ] and voiceless—[ʙ̰]) is similar to that of apical trills. They are produced at the bilabial place of articulation. They are reported to occur at the phonemic level in a few languages such as Piraha and Wari, in South America (Ladefoged and Maddieson, 1996). Figures 12(c) and 12(d) show the spectral characteristics, respectively, of voiced and voiceless bilabial trills. Although it is found that the bilabial trills have highly nonstationary spectral characteristics, a more careful analysis is essential to study the effect of vowel context, and to discriminate them from apical trills. The presence of a voice bar just before the start of trilling can be clearly seen from Fig. 12(c). Figure 13 shows the acoustic cues that can be used for spotting bilabial trills in continuous speech. The utterance contains a voiceless apical trill (discussed in the previous section), followed by a voiced bilabial trill (0.7–1.4 s) and a voiceless bilabial trill (1.4–1.8 s). It can be seen that the zero-frequency analysis brings out prominently the feeble voicing present during the closure region (~0.8 s) in the production of [ʙ]. The excitation strength and the instantaneous F_0 values have much larger fluctuation, compared to apical trills. It is seen from Fig. 13(f) that bilabial trills have a trilling

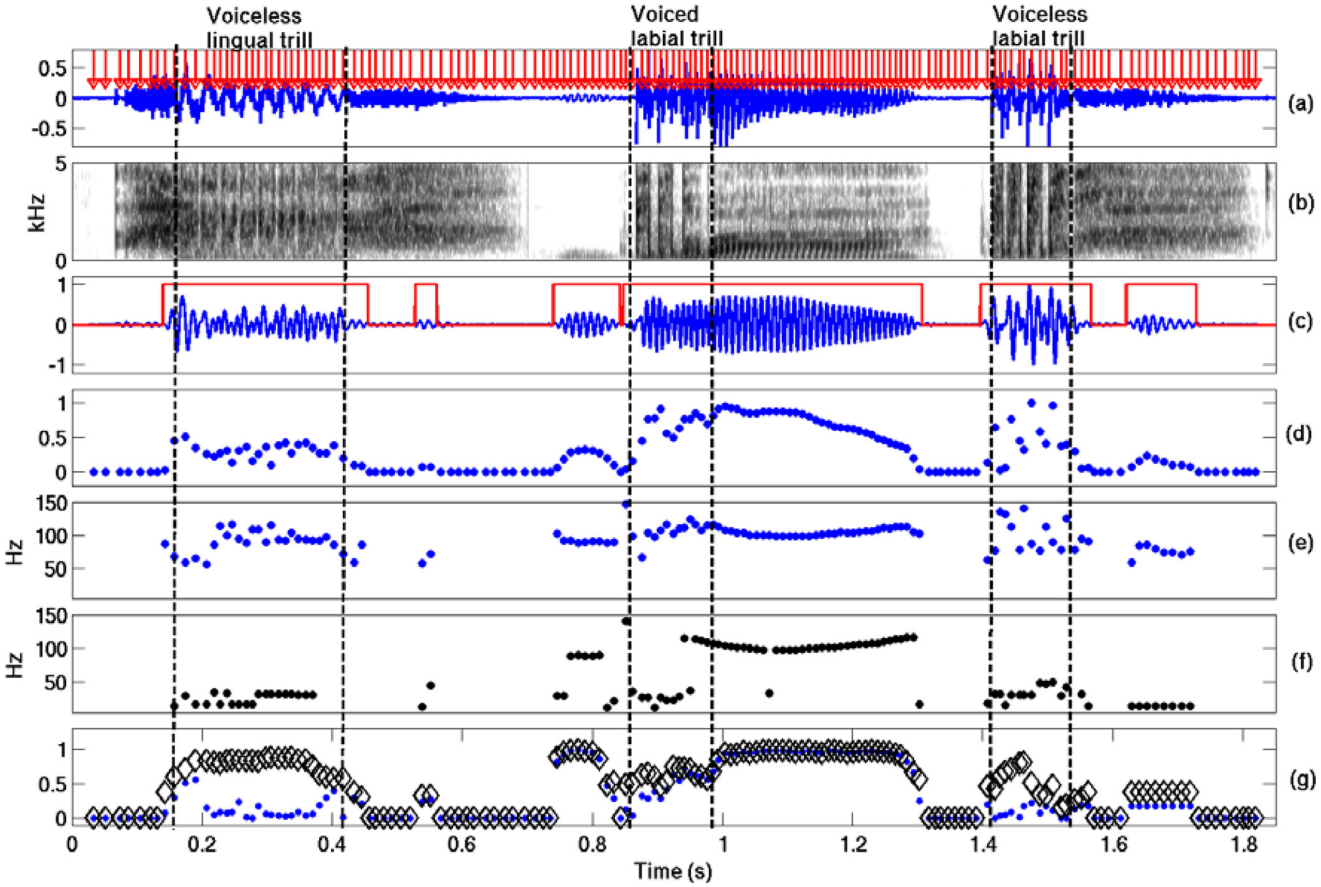


FIG. 13. (Color online) Analysis of voiceless apical trill and labial trills (voiced and voiceless). The utterance shows a VCV unit ([ar̥a]) with a voiceless apical trill (~0.2–0.4 s), followed by two CV units ([ʙaʙa] and ([b̥a b̥a]) containing a voiced labial trill (~0.8–0.9 s) and a voiceless labial trill (~1.4–1.55 s), respectively. (a) Speech waveform along with the glottal epochs, (b) wideband spectrogram, (c) ZFF signal with the voicing region marked, (d) excitation strength, (e) instantaneous F_0, (f) F_{max}—periodicity (in Hz) corresponding to the maximum NCC value ρ_{max}, and (g) NCC values corresponding to trill cycle (diamonds) and glottal cycle (dots).

frequency similar to that of apical trills (~30 Hz). The normalized cross-correlation values have a trend similar to that discussed for apical trills.

VI. SUMMARY AND CONCLUSIONS

Trill sounds, especially the apical trills, are produced with a rapidly changing vocal tract geometry due to the trilling of the apex of the tongue. Although the time-varying vocal tract dynamics results in continuously changing spectral characteristics, it also seems to influence the source of excitation. In this paper the characteristics of voiced apical trills were studied. The excitation source characteristics were studied by examining the features of excitation derived using the zero-frequency analysis of speech signals. The zero-frequency analysis provides information on epoch locations, strength of impulses at epochs and the instantaneous fundamental frequency (F_0). The instantaneous F_0 values vary within a trill cycle, with lower F_0 values in the closed phase of the trill cycle compared to the open phase. This fluctuation could be due to an increase in the pressure gradient across the glottis, when the tapping of the oral cavity by the tongue tip is released. The increase or reduction in the pressure gradient also affects the strengths of the impulses at epochs.

The dynamic spectral characteristics of trills were studied using the zero-time liftering method for analysis of speech. Through this method the instantaneous response of the vocal tract system could be obtained. This new method of analysis enabled us to examine the details of the spectral features during each trill cycle, and also the effect of vowel context on the spectral features of the trills. The spectral characteristics of the trills were examined using the HNGD plots derived using zero-time liftering analysis and group-delay analysis. Together these analysis methods provide good temporal resolution without affecting the spectral features significantly.

The acoustic cues for spotting trills in continuous speech were also explored. The fluctuations of the values of the excitation strength at epochs and the instantaneous F_0 help in discriminating trills from other voiced sounds. Normalized cross-correlation provides an additional cue for spotting trills. These acoustic cues for spotting will be useful if there are steady trill sounds in continuous speech with at least two complete trill cycles. But steady trill sounds with two or more trill cycles are less frequent in continuous speech. The acoustic characteristics of voiceless trills and bilabial trills

were also examined briefly, mainly from the point of view of spotting them in continuous speech.

In summary, this paper demonstrates that trilling in the vocal tract affects the source of excitation. The reduced values of F_0 and the strengths of excitation observed during the closed phase, as compared to the open phase of a trill cycle, are in agreement with the observations made on voiced occlusions (stops). The influence of vowel context on the resonance characteristics of trill was examined using the high temporal resolution of the spectral features provided by the zero-time liftering analysis of speech. Although trills have been reported to have a highly constrained lingual configuration so as to meet the articulatory and aerodynamic requirements, it was shown in this paper that trills do have a significant amount of flexibility in terms of tongue body position, as seen from a large variation in the second formant. The main contribution of this paper is in our ability to extract the dynamic characteristics of both excitation and vocal tract system resonances, using the new signal processing tools like zero-frequency filtering and zero-time liftering.

In this paper only a qualitative description of acoustic features for spotting trills in continuous speech is given. It is important to note that as the vocal tract shape is changing rapidly, and the system is excited mainly due to impulse-like excitation at epochs, the dynamic spectral characteristics sampled near the epoch locations may appear somewhat random during each trill cycle. As the spectral characteristics are changing continuously during the production of trills, and also due to vowel context, it is a challenge to represent them for pattern matching. Further study is needed to develop methods for discriminating different categories of trills such as apical and labial trills. The present study may provide some direction toward such an investigation. It would be interesting to study the contrast between trills and creaky voices, as both produce dynamic characteristics that need to be resolved both in temporal and frequency domain. Although trills are produced with a time-varying vocal tract system excited by a reasonably steady glottal source, creaky voice is produced by exciting a reasonably steady vocal tract by a time-varying glottal source.

Bogert, B. P., Healy, M. J. R., and Tukey, J. W. (**1963**). "The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking," in *Time Series Analysis*, edited by M. Rosenblatt (Wiley, New York), Chap. 15, pp. 209–243.

Catford, J. C. (**1977**). *Fundamental Problems in Phonetics* (Indiana University Press, Bloomington, IN).

Clark, J. E., Yallop, C., and Fletcher, J. (**2007**). *An Introduction to Phonetics and Phonology*, 3rd ed. (Blackwell, Oxford, UK), Chap. 7.

Colantoni, L. (**2006**). "Increasing periodicity to reduce similarity: An acoustic account of deassibilation in rhotics," in *Selected Proceedings of the Second Conference on Laboratory Approaches to Spanish Phonetics and Phonology*, edited by M. Diaz-Campos (Cascadilla Proceedings Project, Somerville, MA), pp. 22–34.

Dhananjaya, N. (**2011**). "Signal processing for excitation-based analysis of acoustic events in speech," Ph.D. thesis, Department of Computer Science and Engineering IIT Madras, Chennai, URL http://speech.iiit.ac.in/svlpubs/phdthesis/dhanu-phd-2011.pdf, (last viewed Jan. 20, 2012).

Dhananjaya, N., Rajendran, S., and Yegnanarayana, B. (**2008**). "Features for automatic detection of voice bars in continuous speech," in *Proceedings of Interspeech*, Brisbane, Australia, pp. 1321–1324.

Dhananjaya, N., and Yegnanarayana, B. (**2010**). "Voiced/nonvoiced detection based on robustness of voiced epochs," IEEE Signal Process. Lett. **17**, 273–276.

Diaz-Campos, M. (**2008**). "Variable production of the trill in spontaneous speech: Sociolinguistic implications," in *Selected Proceedings of the Third Conference on Laboratory Approaches to Spanish Phonology*, edited by L. Colantoni and J. Steele, Cascadilla Proceedings Project, Somerville, MA, pp. 115–127.

Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton & Co., The Hague, The Netherlands), pp. 1–328.

Henriksen, N. C., and Willis, E. W. (**2010**). "Acoustic characterization of phonemic trill production in Jerezano Andalusian Spanish," in *Selected Proceedings of the Fourth Conference on Laboratory Approaches to Spanish Phonology*, edited by M. Ortega-Llebaria, Cascadilla Proceedings Project, Somerville, MA, pp. 115–127.

Herman, I. P. (**2007**). "Sound, speech, and hearing," in *Physics of the Human Body* (Springer, Heidelberg, Germany), pp. 555–619.

Joseph, M. A., Guruprasad, S., and Yegnanarayana, B. (**2006**). "Extracting formants from short segments of speech using group delay functions," in *Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH)*, Pittsburgh, PA, pp. 1009–1012.

Ladefoged, P., Cochran, A., and Disner, S. F. (**1977**). "Laterals and trills," J. Int. Phonet. Assoc. **7**, 46–54.

Ladefoged, P., and Maddieson, I. (**1996**). *Sounds of World's Languages* (Blackwell, Oxford, UK), Chap. 7.

Lindau, M. (**1985**). "The story of /r/," in *Phonetic Linguistics: Essays in Honor of P. Ladefoged*, edited by V. Fromkin (Academic, Orlando, FL), pp. 157–167.

Lipski, J. (**1994**). *Latin American Spanish* (Longman Linguistics Library, New York), pp. 1–440.

Maddieson, I. (**1984**). *Patterns of sounds* (Cambridge University Press, Cambridge, UK), pp. 1–422.

McGowan, R. S. (**1992**). "Tongue-tip trills and vocal-tract wall compliance," J. Acoust. Soc. Am. **91**, 2903–2910.

Murty, K. S. R., and Yegnanarayana, B. (**2008**). "Epoch extraction from speech signals," IEEE Trans. Audio, Speech, Lang. Process. **16**, 1602–1613.

Murty, K. S. R., Yegnanarayana, B., and Joseph, M. A. (**2009**). "Characterization of glottal activity from speech signals," IEEE Signal Process. Lett. **16**, 469–472.

Oppenheim, A. V., and Schafer, R. W. (**1975**). *Digital Signal Processing* (Prentice–Hall, Englewood Cliffs, NJ), pp. 1–585.

Recasens, D. (**1991**). "On the production characteristics of apicoalveolar taps and trills," J. Phonet. **19**, 267–280.

Recasens, D., and Pallars, M. D. (**1999**). "A study of/r/and/r/in the light of the DAC coarticulation model," J. Phonet. **27**, 143–169.

Ruhlen, M. (**1987**). *A Guide to the World's Languages* (Stanford University Press, Palo Alto, CA), Vol. 1, pp. 1–492.

Solé, M. J. (**2002**). "Aerodynamic characteristics of trills and phonological patterning," J. Phonet. **30**, 655–688.

Spajic, S., Ladefoged, P., and Bhaskararao, P. (**1996**). "The trills of Toda," J. Int. Phonet. Assoc. **26**, 1–21.

Stevens, K. N. (**1999**). *Acoustic Phonetics* (The MIT Press, Cambridge, MA), Chap. 2.

van den Berg, J. (**1957**). "Subglottic pressures and vibrations of the vocal folds," Folia Phoniat. **9**, 65–71.

Westbury, J. R. (**1983**). "Enlargement of the supraglottal cavity and its relation to stop consonant voicing," J. Acoust. Soc. Am. **73**, 1322–1336.

Yegnanarayana, B. (**1978**). "Formant extraction from linear prediction phase spectra," J. Acoust. Soc. Am. **63**, 1638–1640.

Yegnanarayana, B., and Murthy, H. A. (**1992**). "Significance of group delay functions in spectrum estimation," IEEE Trans. Signal Process. **40**, 2281–2289.

Yegnanarayana, B., and Murty, K. S. R. (**2009**). "Event-based instantaneous fundamental frequency estimation from speech signals," IEEE Trans. Audio, Speech, Lang. Process. **17**, 614–624.

2

IIT Kanpur

Progress Report of IIT Kanpur

A. General

A.1	Name of the Project	:	**Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages**
	Our Reference Letter No	:	
A.2	Executing Agency	:	IIT Kanpur
A.3	Chief Investigator with	:	Dr. Rajesh Hegde
	Designation	:	Associate Professor
	Co-Chief Investigators with	:	Dr. Harish Karnick, Professor
	Designation	:	
A.4	Project staffs with	:	(1)Gaurav Kumar Singh (MTech)
	Qualification	:	(2) Ishtiyaq Husain (M.C.A)
		:	(3) Sukhjeet Kaur (M.C.A)
		:	(4) Laxmi Panday (B.Tech)
		:	(5) Sumit Tiwari (MCA)

A.5 Total Cost of the Project as approved by DIT

(i) Original	:	Rs. 40.25 Lakhs
(ii) Revised, if any	:	

A.6	Project Sanction Date	:	23-12-2011
A.7	Date of Completion	:	Not Applicable
	(i) Original	:	
	(ii) Revised, if any	:	
A.8	Date on which last progress report was Submitted	:	28-02-2014

A.9 Work in Progress (Details are given in technical report in Appendix 2.1)

Appendix 2.1

Detailed Technical Report of IIT Kanpur

Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages Hindi Language

Report of IIT Kanpur

1. Database collection and transcription

1.1 Data collection:

Data has been downloaded for three different categories from the Internet 20 hours of speech data has been collected in which 10 hours of data is collected for transcription and 10 hours of data is collected as raw data.

• **Read speech:** Direct context data has been collected from newsonair.nic.in in mp3 format and converted into the MS-wave format with sampling rate of 16 KHz and bit rate of 16-bit PCM.

• **Lecture speech:** Extempore context data has been collected from video sharing websites like YouTube in a multimedia format and converted into the MS-wav format with sampling rate of 16 KHz and bit rate of 16-bit PCM.

• **Conversational speech:** Casual context has been collected from video sharing websites like YouTube in multimedia format and converted into the MS-wav format with sampling rate of 16 KHz and bit rate of 16-bit PCM.

1.2 Test beds for Data Collection:

Currently data is collected from the Internet using a smart Internet enabled television and desktops. The setup is shown in Figure 1.

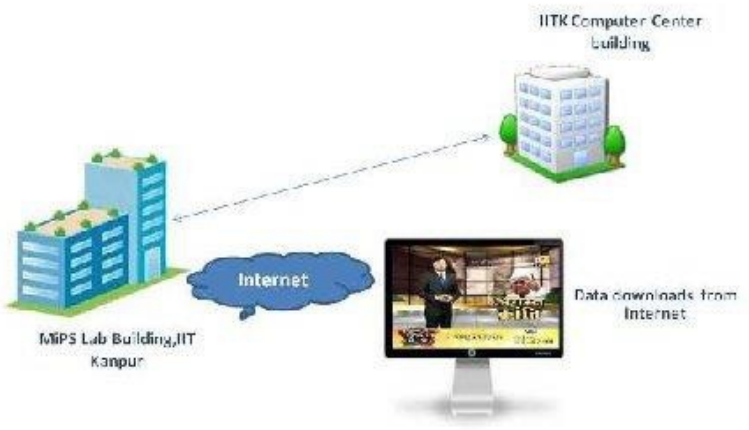


Figure 1: Data Collection Process

2. Data Transcription

Phonetic transcription uses phonetic symbols for the representation of speech sounds. The information that needs to be captured in the transcription is essentially what the speaker has spoken and not what the speaker intended to speak. The choice of symbol must be such that it captures all the phonetic variations in speech. Existing phonetic engines for Indian languages use syllable like units as sub words units.

2.1 Transcription using ARPAbet :

Transcription has been done using Standard English phones using ARPAbet symbols. This phoneme (or more accurately, phone) set is based on the ARPAbet symbol set developed for speech recognition uses. The phone set is the list of 'phones', or speech sounds, that the engine can recognize. While building acoustic models and pronunciations for words, phone set can be made to use any set of units. The acoustic models searches for the speech sounds (phones), and the word pronunciations are also given in terms of the phones in the phone set. The example illustrating the transcription with standard phones is shown in Table I

Name	Transcription using ARPAbet
Allahabad	ah l aa hh aa b aa d
Gorakhpur	g ow r ah k p uh r
Agra	ae g r ah
Farrukhabad	f aa r uw k aa b aa dh

TABLE I: Table illustrating transcription using ARPAbet

2.2 Transcription using IPA symbols:

The data is segmented into reasonably shorter duration for better verification, and efficient transcription. During transcription, the speech signal is listened carefully and Transcription performed for minimum transcription error. Transcription has been done using the International Phonetic Alphabet (IPA) chart as shown in Table II

Name	Transcription using IPA symbols
Allahabad	ala:haba:d
Gorakhpur	gorək^{h}pur
Agra	a:gra:
Farrukhabad	fərruk^{h}a:ba:d

TABLE II: Table illustrating transcription using IPA

Few More Examples of IPA Transcription:

Sentence – रावण से घमासान युद्ध में उसकी नाभि में अमृत का भेद ज्ञात होने पर

IPA – ra:ʋəɳsegʰəma:sa:njuddʰmẽuski:na:bʰi:meamɾitka:bʰedgja:thonepəɾ

Sentence – अर्थात रावण की प्रकृति वाला व्यक्ति कभी भी सफल नहीं हो सकता है

IPA – artʰa:ɾtra:ʋəɳki:prəkɾəti:ʋa:la:vjəkti:kəbʰi:bʰi:səfəlnəhi:hosəkta:hɛ

Sentence	Transcription at syllable level
namaskar apka swagat hai	nəm əs ka:r a:p ka: sva: gət hɛ
akashvani se prastut hai	a: ka:ʃ va: ni: se prəs tut hɛ
aam admi ki sarkar hai	a:m a:d mi: ki: səɾ ka:r hɛ
aj charcha ka vishay hai	a:ʤ tʃəɾ tʃa: ka: vi ʃəj hɛ

2.3 Details of Transcription:

Audio data must be collected according to specific domain (parliament speech, news) that contains all vowels and consonants of Hindi. Data has been collected from video sharing websites like YouTube. The audio data is collected in uncompressed format with a sampling rate of 8 KHz. This data set is used in bootstrapping the audio search system to improve recognition accuracy. The fillers used in the transcription of data has been listed in Table2.

Fillers	**Explanations**
<bgnoise>	Background noise
<chnoise>	Channel Noise
<ah>	Clearing of throat
<fstart>	False Start

Table 2: Fillers used in the transcription of data

2.2 Commonly Occurring Monophones in the entire vocabulary

The most commonly occurring monophones in the entire vocab are listed in Table 3.

Monophones	@	a	e	i	k
Percentage Occ.	11.16	9.59	8.65	6.92	6.29
Monophones	R	m	u	n	r
Percentage Occ.	3.93	3.77	3.46	3.46	3.46

Table 3: Frequently occurring monophones with their percentage occurrences

2.3 Parameters used in acoustic model building

The Parameters used for Acoustic model building are listed in Table 4.

Parameters	**Hindi Database**
Lower frequency	200
Upper Frequency	3400
No of tied states	969
No of mixtures	16
Language Model type	Trigram
Tools to check accuracy	Word Align

Table 4: Acoustic model parameters

3. Prosodic Labeling

Prosodic features have no direct correspondence to written characters of sentence and are unique to spoken language. Prosody refers to the suprasegmental features of natural speech (such as rhythm and intonation) that are used to convey linguistic and paralinguistic information (such as emphasis, intention, attitude and emotion). Steps for prosodic labeling are as follows:

- Syllabification: Transcribed data is segmented at syllable boundaries.
- Pitch Marking: Perform pitch marking according to the pitch variations in the audio wave file. While marking, consider only isolated segments. Mark the signal with High (H), Low (L), Very High (VH), and Very Low (VL). Pitch contours not considered here for marking.
- Break Index: Mark the signal with hypothetical Break (B0), physical break (B1, B2, B3). B1, B2, B3 can be varied according to the length of the breaks between the segments. The syllables which are lying in particular pitch contour are marked with the same labels as defined for that particular pitch contour.

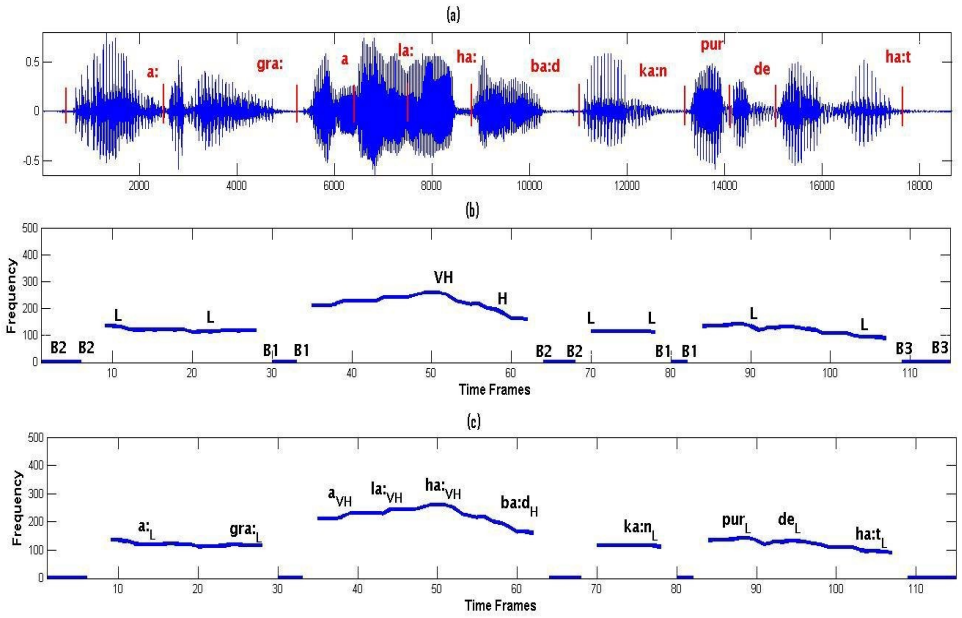
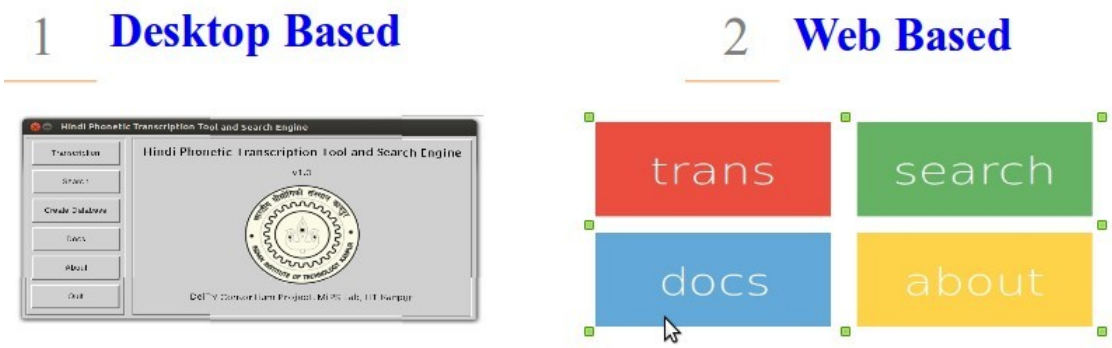


Figure 2: Prosodic labeling an example

4. Development of Phonetic Transcription Tool

Phonetic engine uses the acoustic phonetic information present in the speech signal and convert into the data in some symbolic form. The conversion from speech signal to the appropriate symbolic form requires efficient transcription for good speech recognition performance. The transcription is useful in the systematic representation of language in written form. The information that needs to be captured in the transcription is essentially what the speaker has spoken and not what the speaker intended to speak. The choice of symbol must be such that it captures all the phonetic variations in speech. Existing phonetic engine for Indian languages uses syllable like units as sub words units. Here we are using a sequence of symbols based on International Phonetic Alphabet (IPA) as the sub word units. The IPA is designed to represent the qualities of speech that are distinctive in oral language: Phonemes, intonation and the separation of words and syllables. IPA symbols are composed of one or more elements of two basic types, letters and diacritics. In the current transcription, a sequence of International Phonetic Alphabet (IPA) are used as sub word units since IPA provides one symbol for each distinctive sound.



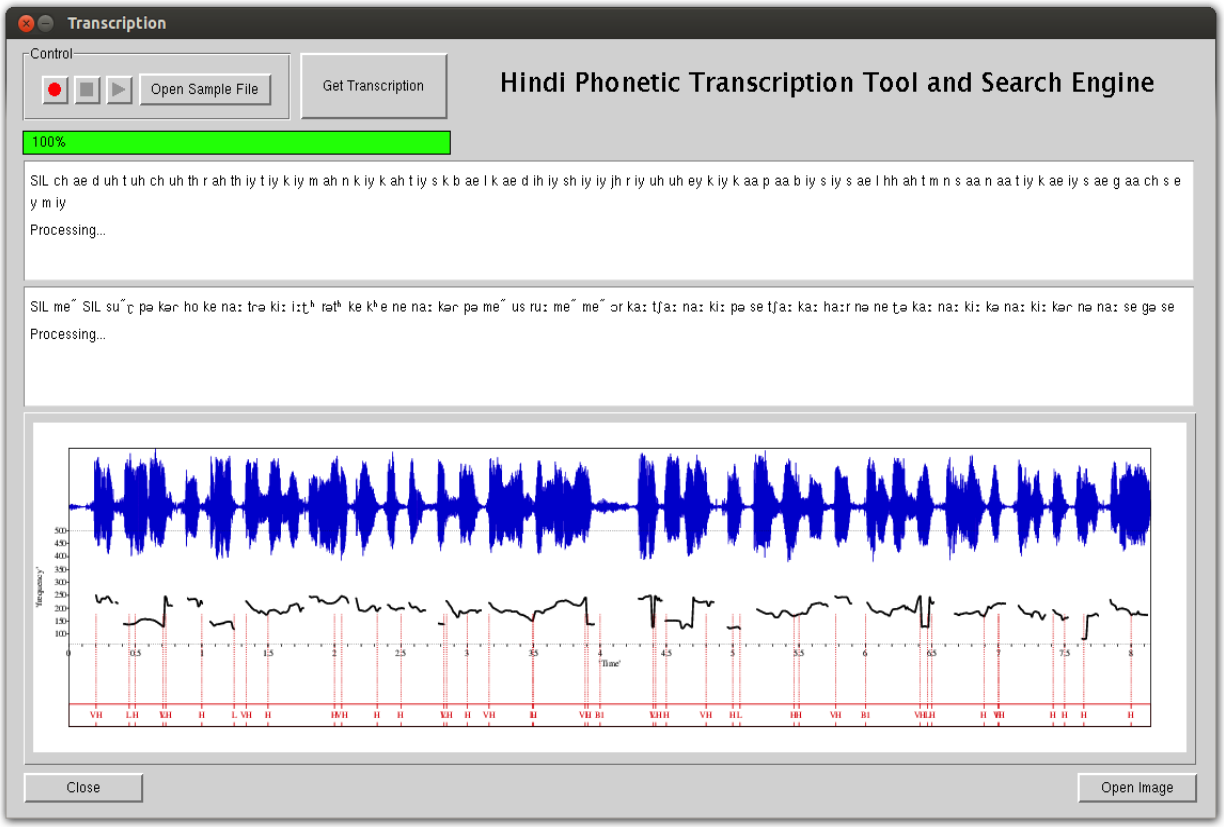


Figure 3: Transcription of a Sentence

5. Development of Audio Search Tool

Speech search engine is a software program that is used to search audio occurrences of spoken words or phrases by using a sequence of symbol produces by phonetic engine as input. Speech based retrieval systems deal with searching and retrieving spoken documents in response of spoken queries. Spoken documents are the speech files that are converted to phonetic sequence by phonetic engine. Phonetic searching comprises two phases-indexing and searching. The first phase indexes
The input speech to produce a phonetic search track and is performed by once. The second phase performed whenever a search is needed for a word or phrase, is searching the phonetic speech tracks which are stored with reference to the spoken documents. Keywords spotting can also be used which deals with the identification of keywords in utterances. The keywords are extracted from spoken query using phonetic dictionary. The documents are retrieved to the user on the basis of keyword occurrences. In the subsequent sections we describe an audio search system developed at IIT Kanpur.

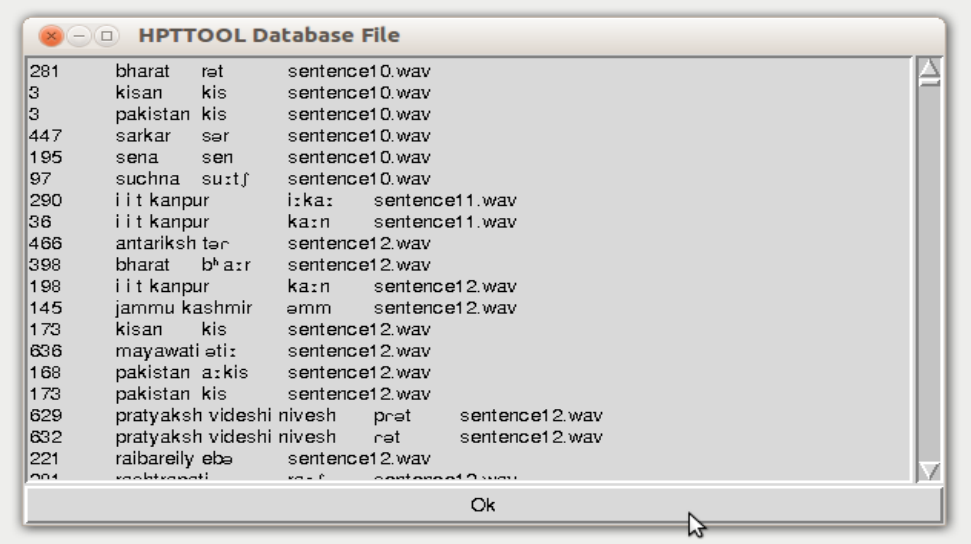


Figure 4: Databases File

5.1 Audio search Tool

In this system, the test audio file is automatically transcribed to get word level output of keyword that needs to be searched in the database. The complete audio search system is illustrated in Figure 5.

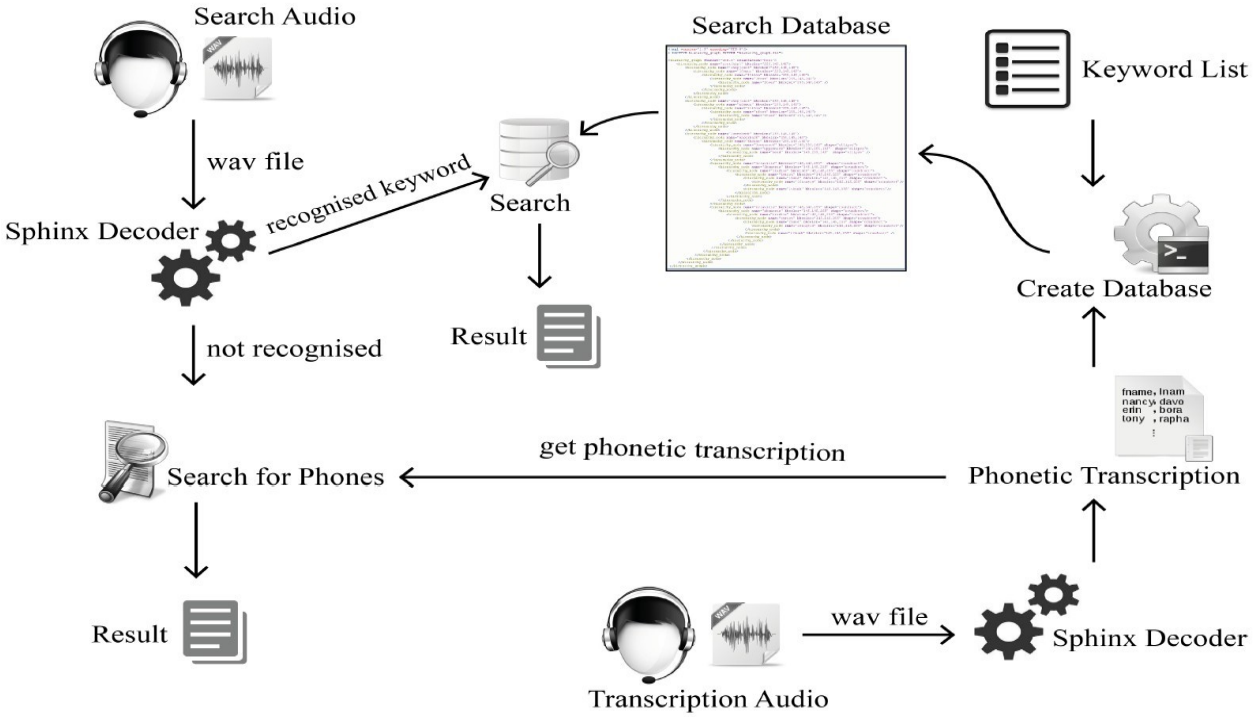


Figure 5: Complete audio search system

If the keyword and its phonetic transcription are already present in the database then this keyword is searched in the database to get the time stamp of the keyword. If keyword is an out of vocabulary word, then its phonetic transcription is automatically derived and the time stamp corresponding to each phone is also derived. In order to get the time stamp of keyword, an automatically annotated database is created which consists of text file containing keyword, their phonetic transcription with all possible combinations of at least three consecutive phones along with their time stamp. Subsequently every phone of keyword is matched with the recognized output sequence of phones as shown in Figure 3. If at least three consecutive phones in the overall string match then the possibility of the keyword existing in that particular time frame is very high. A pictorial illustration of the sliding phone protocol is illustrated in Figure 6 for the particular word 'Rashtrapati'. It can be noted that keyword 'Rashtrapati' is enclosed in boxes. The phonetic transcription below the actual boxed transcription is that of a long test sentence which consists of the keyword. It may be noted from Figure 6, that in a particular case the sliding phone protocol is able to detect the presence of the keyword in a long test utterance.

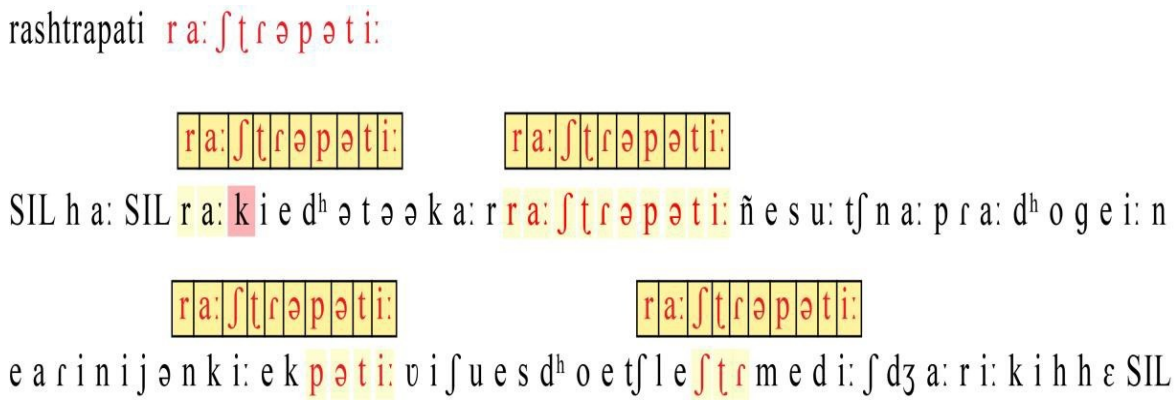


Figure 6: Sliding phone protocol for audio search

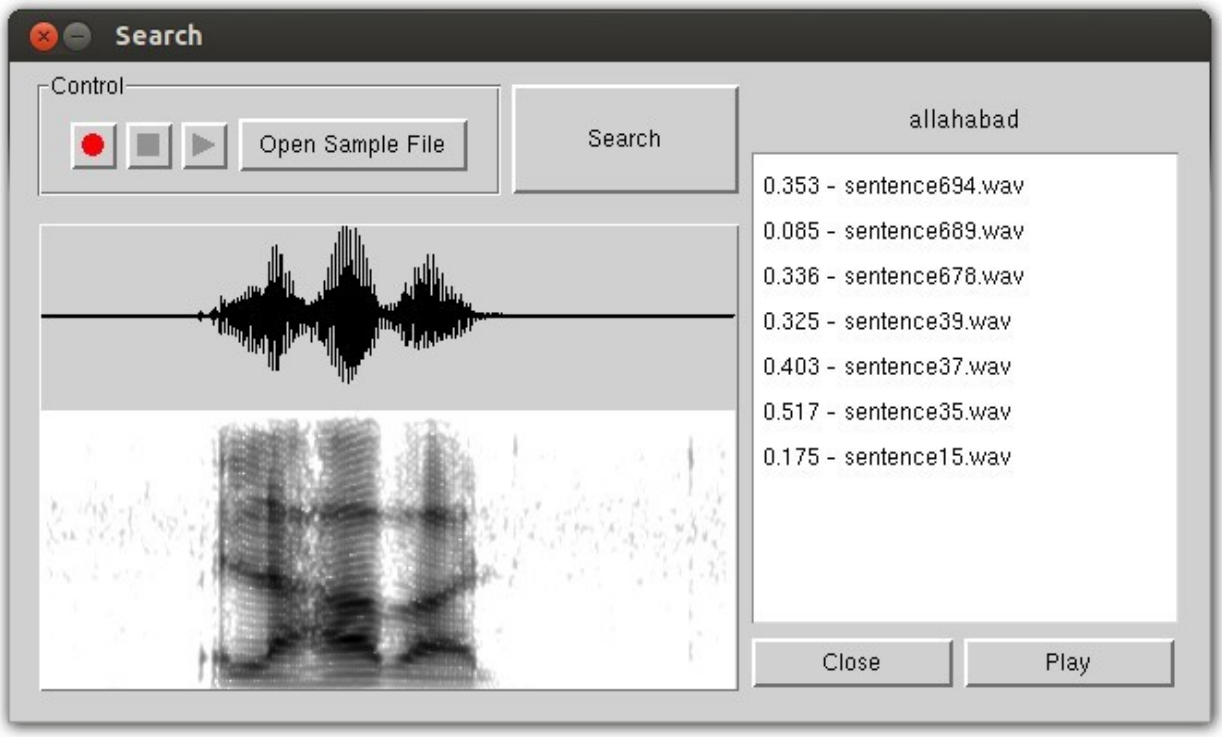


Figure 7: shows Search of a Sentence

6. Performance Evaluation of Audio Search System:

In order to train acoustic models and test the system, a corpus suitable for training of recognition system that contains high quality transcription is needed. The data must be collected in such a way that they are rich in phones to train HMMs. Model has been tested using keywords that occur most frequently and recognition accuracy turns out to be 92.30 table shows performance evolution of Transcription & Search

Transcription Performance

Keyword：

No. of Speakers	No. of Test Keywords	Phone Recognition Accuracy
5	30	84.18%

Sentence：

No. of Speakers	No. of Test Sentences	Phone Recognition Accuracy
3	50	32.36%

Search Performance

No. of Speakers	No. of Test keywords	Phone recognition
25	450	51.24%

7. Summary & Future work

In this work, a framework is presented for indexing speech databases. The pitch accents are first assigned at the pitch breaks and then mapped to the syllables falling in that duration. Subsequently models are trained for most frequently occurring syllables with phonetic tags. An automated time stamping system and sliding phone protocol is used for searching the keyword in the development of a phone based audio indexing system .This protocol helps in reasonably improving the speed of the phone based recognition system. Future work will focus on developing fast methods for domain independent audio search and indexing.

8. Appendix

8.1 Detailed Evaluation for Keywords

Sample Wav

Keywords	Output
Ajodhya	अयोध्या *** A J o d^h *** Y a□ *** (AYODHYA) SIL A□ Dʒ o d^h Dʒ R a□ SIL (AYODHYA) Words: 6 Correct: 3 Errors: 6 Percent correct = 50.00% Error = 100.00% Accuracy = 0.00% **Insertions: 3 Deletions: 0 Substitutions: 3**
Bank	बैंक *** *** B □̃ *** *** k *** (BANK) SIL E M □̃ ə M k SIL (BANK) Words: 3 Correct: 2 Errors: 6 Percent correct = 66.67% Error = 200.00% Accuracy = -100.00% **Insertions: 5 Deletions: 0 Substitutions: 1**
Bharat	भारत *** b^h a□ r ə t (BHARAT) SIL b^h a□ r ə t (BHARAT) Words: 5 Correct: 5 Errors: 1 Percent correct = 100.00% Error = 20.00% Accuracy = 80.00% **Insertions: 1 Deletions: 0 Substitutions: 0**
Bhopal	भोपाल *** *** *** *** B^h O p *** A□ L (BHOPAL) SIL SIL SIL H D M p □̃ D A□ (BHOPAL) Words: 5 Correct: 1 Errors: 9 Percent correct = 20.00% Error = 180.00% Accuracy = -80.00% **Insertions: 5 Deletions: 0 Substitutions: 4**
Cbi	सी बी आई *** S i□ b *** i□ a□ i□ *** (CBI) SIL Tʃ i□ bd̪ i□ a□ SIL (CBI) Words: 6 Correct: 5 Errors: 4 Percent correct = 83.33% Error = 66.67% Accuracy = 33.33% **Insertions: 3 Deletions: 0 Substitutions: 1**
Cheen	चीन *** tʃ i□ n (CHEEN) SIL tʃ i□ n (CHEEN) Words: 3 Correct: 3 Errors: 1 Percent correct = 100.00% Error = 33.33% Accuracy = 66.67% **Insertions: 1 Deletions: 0 Substitutions: 0**
Cricket	क्रिकेट *** k ɾ i k e t *** (CRICKET) SIL k ɾ i k e t SIL (CRICKET) Words: 6 Correct: 6 Errors: 2 Percent correct = 100.00% Error = 33.33% Accuracy = 66.67% **Insertions: 2 Deletions: 0 Substitutions: 0**
Koyla gate ghotale	कोयला गेट घोटाले *** k o j ə l a□ g e t g^h o t a□ l e *** (KOYLA_GATE_GHOTALE) SIL k o j ə l a□ g e t g^h o t a□ l e SIL (KOYLA_GATE_GHOTALE)

	Words: 15 Correct: 15 Errors: 2 Percent correct = 100.00% Error = 13.33% Accuracy = 86.67% **Insertions: 2 Deletions: 0 Substitutions: 0**
Manmohan singh	मनमोहन सिंह *** m ə n m o h ə n s ĩ G^{h} (MANMOHAN_SINGH) SIL m ə n m o h ə n s ĩ J (MANMOHAN_SINGH) Words: 11 Correct: 10 Errors: 2 Percent correct = 90.91% Error = 18.18% Accuracy = 81.82% **Insertions: 1 Deletions: 0 Substitutions: 1**
Mayawati	मायावती *** m a j a ə t i *** (MAYAWATI) SIL m a j a ə t i SIL (MAYAWATI) Words: 8 Correct: 8 Errors: 2 Percent correct = 100.00% Error = 25.00% Accuracy = 75.00% **Insertions: 2 Deletions: 0 Substitutions: 0**
Rashtrapati	राष्ट्रपति *** r a ʃ ʈ ɾ ə P ə t i (RASHTRAPATI) SIL r a ʃ ʈ ɾ *** *** ə t i (RASHTRAPATI) Words: 10 Correct: 8 Errors: 3 Percent correct = 80.00% Error = 30.00% Accuracy = 70.00% **Insertions: 1 Deletions: 2 Substitutions: 0**
Videshi pratyaksh nivesh	विदेशी प्रत्यक्ष निवेश *** ʊ i d e ʃ i p ɾ ə t j ə k ʃ n i ʊ e ʃ (VIDESHI_PRATYAKSH_NIVESH) SIL ʊ i d e ʃ i p ɾ ə t j ə k ʃ n i ʊ e ʃ (VIDESHI_PRATYAKSH_NIVESH) Words: 19 Correct: 19 Errors: 1 Percent correct = 100.00% Error = 5.26% Accuracy = 94.74% **Insertions: 1 Deletions: 0 Substitutions: 0**
Vishvidyalay	विश्वविद्यालय *** ʊ i ʃ ʊ ə ʊ i d j a l ə j (VISHVAVIDYALAY) SIL ʊ i T ʊ ə ʊ i d j a l ə j (VISHVAVIDYALAY) Words: 13 Correct: 12 Errors: 2 Percent correct = 92.31% Error = 15.38% Accuracy = 84.62% **Insertions: 1 Deletions: 0 Substitutions: 1**

LIVE

Keywords	Output
Allahabad	अलाहाबाद *** *** a l a् h a् b a् d *** *** *** (ALLAHA) SIL A a l a् h a् b a् d SIL SIL SIL (ALLAHA) Words: 8 Correct: 8 Errors: 5 Percent correct = 100.00% Error = 62.50% Accuracy = 37.50% **Insertions: 5 Deletions: 0 Substitutions: 0**
America	अमेरिका *** *** *** a m e r i k a् *** (AMERICA) SIL K SIL a m e r i k a् SIL (AMERICA) Words: 7 Correct: 7 Errors: 4 Percent correct = 100.00% Error = 57.14% Accuracy = 42.86% **Insertions: 4 Deletions: 0 Substitutions: 0**
Anna Hazare	अन्ना हजारे *** A n n a् H ə DZ A् R E (ANNA) SIL S n n a् *** ɜ̃ S ə G^h SIL (ANNA) Words: 10 Correct: 3 Errors: 8 Percent correct = 30.00% Error = 80.00% Accuracy = 20.00% **Insertions: 1 Deletions: 1 Substitutions: 6**
Bharat	भारत *** *** b^h a् r ə *** *** T (BHARAT) SIL E b^h a् r ə D^h D SIL (BHARAT) Words: 5 Correct: 4 Errors: 5 Percent correct = 80.00% Error = 100.00% Accuracy = 0.00% **Insertions: 4 Deletions: 0 Substitutions: 1**
Bhopal	भोपाल *** *** *** *** *** *** b^h o p *** *** *** *** *** *** A् L (BHOPAL) SIL S K ʃ SIL P b^h o p B ə ɾ S SIL S I SIL (BHOPAL) Words: 5 Correct: 3 Errors: 14 Percent correct = 60.00% Error = 280.00% Accuracy = -180.00% **Insertions: 12 Deletions: 0 Substitutions: 2**
CBI	सी बी आई *** *** *** *** *** *** S i् b i् *** *** A् I् (CBI) SIL E O D ɾ ə Tʃ i् b i् N SIL G L (CBI) Words: 6 Correct: 3 Errors: 11 Percent correct = 50.00% Error = 183.33% Accuracy = -83.33% **Insertions: 8 Deletions: 0 Substitutions: 3**
Cheen	चीन *** *** *** *** tʃ *** *** *** I् N (CHEEN) SIL D K ʃ tʃ N A् ʊ SIL SIL (CHEEN) Words: 3 Correct: 1 Errors: 9 Percent correct = 33.33% Error = 300.00% Accuracy = -200.00% **Insertions: 7 Deletions: 0 Substitutions: 2**
Cricket	क्रिकेट *** *** *** *** K ɾ i k e *** *** ʈ (CRICKET) SIL S T N SIL K i k e ʊ D SIL (CRICKET) Words: 6 Correct: 3 Errors: 9 Percent correct = 50.00% Error = 150.00% Accuracy = -50.00% **Insertions: 6 Deletions: 0 Substitutions: 3**

Green House Gases	ग्रीन हाउस गैसेस G ɾ i N H a *** U S g s õ *** *** *** (GHG) SIL SIL i G ʊ a S t ɾ g ɛ s õ SIL SIL SIL (GHG) Words: 12 Correct: 6 Errors: 10 Percent correct = 50.00% Error = 83.33% Accuracy = 16.67% **Insertions: 4 Deletions: 0 Substitutions: 6**
Vishvidyalay	विश्वविद्यालय *** *** ʊ I ʃ ʊ ə ʊ i D J a I ***ə J (VISHVAVIDYALAY) SIL E ʊ J E ʃ N G i *** K a I E ə SIL (VISHVAVIDYALAY) Words: 13 Correct: 5 Errors: 11 Percent correct = 38.46% Error = 84.62% Accuracy = 15.38% **Insertions: 3 Deletions: 1 Substitutions: 7**
Videshi pratyaksh nivesh	विदेशी प्रत्यक्ष निवेश ʊ I D E ʃ I p ɾ ə t J ə k ʃ n i ʊ e ʃ *** *** *** *** *** (VPN) SIL ʊ I T ʃ ə ɾ p ɾ *** t T ə k ʃ n i ʊ e ʃ SIL ʈ T A SIL (VPN) Words: 19 Correct: 11 Errors: 13 Percent correct = 57.89% Error = 68.42% Accuracy = 31.58% **Insertions: 5 Deletions: 1 Substitutions: 7**
Varsha	वर्षा *** *** *** *** *** ʊ ə ɾ ʃ a *** *** *** *** *** *** *** (VARSHA) SIL G M B ə ʊ I D J a Õ SIL D SIL ʊ T SIL (VARSHA) Words: 5 Correct: 2 Errors: 15 Percent correct = 40.00% Error = 300.00% Accuracy = -200.00% **Insertions: 12 Deletions: 0 Substitutions: 3**

8.2: Detailed Evaluation for Sentences

Sample wav

sentence1	Words: 118 Correct: 27 Errors: 92 Percent correct = 22.88% Error = 77.97% Accuracy = 22.03% **Insertions: 1 Deletions: 35 Substitutions: 56**
sentence2	Words: 91 Correct: 20 Errors: 71 Percent correct = 21.98% Error = 78.02% Accuracy = 21.98% **Insertions: 0 Deletions: 22 Substitutions: 49**
sentence3	Words: 94 Correct: 20 Errors: 75 Percent correct = 21.28% Error = 79.79% Accuracy = 20.21% **Insertions: 1 Deletions: 28 Substitutions: 46**
sentence4	Words: 151 Correct: 37 Errors: 119 Percent correct = 24.50% Error = 78.81% Accuracy = 21.19% **Insertions: 5 Deletions: 37 Substitutions: 77**
sentence5	Words: 89 Correct: 19 Errors: 71 Percent correct = 21.35% Error = 79.78% Accuracy = 20.22% **Insertions: 1 Deletions: 29 Substitutions: 41**
sentence6	Words: 158 Correct: 30 Errors: 128 Percent correct = 18.99% Error = 81.01% Accuracy = 18.99% **Insertions: 0 Deletions: 35 Substitutions: 93**
sentence7	Words: 147 Correct: 32 Errors: 116 Percent correct = 21.77% Error = 78.91% Accuracy = 21.09% **Insertions: 1 Deletions: 44 Substitutions: 71**
sentence8	Words: 51 Correct: 11 Errors: 42 Percent correct = 21.57% Error = 82.35% Accuracy = 17.65% **Insertions: 2 Deletions: 18 Substitutions: 22**
sentence9	Words: 70 Correct: 14 Errors: 56 Percent correct = 20.00% Error = 80.00% Accuracy = 20.00% **Insertions: 0 Deletions: 13 Substitutions: 43**
sentence10	Words: 77 Correct: 11 Errors: 66 Percent correct = 14.29% Error = 85.71% Accuracy = 14.29% **Insertions: 0 Deletions: 20 Substitutions: 46**
sentence11	Words: 85 Correct: 18 Errors: 68 Percent correct = 21.18% Error = 80.00% Accuracy = 20.00% **Insertions: 1 Deletions: 17 Substitutions: 50**
sentence12	Words: 55 Correct: 13 Errors: 42 Percent correct = 23.64% Error = 76.36% Accuracy = 23.64% **Insertions: 0 Deletions: 13 Substitutions: 29**
sentence13	Words: 43 Correct: 4 Errors: 39 Percent correct = 9.30% Error = 90.70% Accuracy = 9.30% **Insertions: 0 Deletions: 7 Substitutions: 32**
sentence14	Words: 42 Correct: 6 Errors: 36 Percent correct = 14.29% Error = 85.71% Accuracy = 14.29% **Insertions: 0 Deletions: 7 Substitutions: 29**
sentence15	Words: 51 Correct: 11 Errors: 41 Percent correct = 21.57% Error = 80.39% Accuracy = 19.61% **Insertions: 1 Deletions: 13 Substitutions: 27**
sentence16	Words: 104 Correct: 23 Errors: 83 Percent correct = 22.12% Error = 79.81% Accuracy = 20.19% **Insertions: 2 Deletions: 32 Substitutions: 49**
sentence17	Words: 116 Correct: 21 Errors: 95 Percent correct = 18.10% Error = 81.90% Accuracy = 18.10%

	Insertions: 0 Deletions: 26 Substitutions: 69
sentence18	Words: 66 Correct: 11 Errors: 55 Percent correct = 16.67% Error = 83.33% Accuracy = 16.67% **Insertions: 0 Deletions: 20 Substitutions: 35**
sentence19	Words: 66 Correct: 19 Errors: 49 Percent correct = 28.79% Error = 74.24% Accuracy = 25.76% **Insertions: 2 Deletions: 20 Substitutions: 27**
sentence20	Words: 124 Correct: 30 Errors: 95 Percent correct = 24.19% Error = 76.61% Accuracy = 23.39% **Insertions: 1 Deletions: 42 Substitutions: 52**
sentence21	Words: 151 Correct: 30 Errors: 121 Percent correct = 19.87% Error = 80.13% Accuracy = 19.87% **Insertions: 0 Deletions: 62 Substitutions: 59**
sentence22	Words: 85 Correct: 18 Errors: 67 Percent correct = 21.18% Error = 78.82% Accuracy = 21.18% **Insertions: 0 Deletions: 25 Substitutions: 42**
sentence23	Words: 111 Correct: 30 Errors: 88 Percent correct = 27.03% Error = 79.28% Accuracy = 20.72% **Insertions: 7 Deletions: 23 Substitutions: 58**
sentence24	Words: 111 Correct: 17 Errors: 94 Percent correct = 15.32% Error = 84.68% Accuracy = 15.32% **Insertions: 0 Deletions: 29 Substitutions: 65**
sentence25	Words: 127 Correct: 25 Errors: 102 Percent correct = 19.69% Error = 80.31% Accuracy = 19.69% **Insertions: 0 Deletions: 49 Substitutions: 53**
sentence26	Words: 84 Correct: 18 Errors: 67 Percent correct = 21.43% Error = 79.76% Accuracy = 20.24% **Insertions: 1 Deletions: 25 Substitutions: 41**
sentence27	Words: 130 Correct: 28 Errors: 102 Percent correct = 21.54% Error = 78.46% Accuracy = 21.54% **Insertions: 0 Deletions: 41 Substitutions: 61**
sentence28	Words: 69 Correct: 10 Errors: 60 Percent correct = 14.49% Error = 86.96% Accuracy = 13.04% **Insertions: 1 Deletions: 28 Substitutions: 31**
sentence29	Words: 65 Correct: 14 Errors: 55 Percent correct = 21.54% Error = 84.62% Accuracy = 15.38% **Insertions: 4 Deletions: 7 Substitutions: 44**
sentence30	Words: 74 Correct: 12 Errors: 63 Percent correct = 16.22% Error = 85.14% Accuracy = 14.86% **Insertions: 1 Deletions: 32 Substitutions: 30**
sentence31	Words: 87 Correct: 19 Errors: 68 Percent correct = 21.84% Error = 78.16% Accuracy = 21.84% **Insertions: 0 Deletions: 32 Substitutions: 36**
sentence32	Words: 71 Correct: 11 Errors: 60 Percent correct = 15.49% Error = 84.51% Accuracy = 15.49% **Insertions: 0 Deletions: 17 Substitutions: 43**
sentence33	Words: 45 Correct: 12 Errors: 33 Percent correct = 26.67% Error = 73.33% Accuracy = 26.67% **Insertions: 0 Deletions: 11 Substitutions: 22**
sentence34	Words: 48 Correct: 10 Errors: 39 Percent correct = 20.83% Error = 81.25% Accuracy = 18.75% **Insertions: 1 Deletions: 15 Substitutions: 23**
sentence35	Words: 74 Correct: 11 Errors: 65 Percent correct = 14.86% Error = 87.84% Accuracy = 12.16% **Insertions: 2 Deletions: 7 Substitutions: 56**
sentence36	Words: 96 Correct: 21 Errors: 75 Percent correct = 21.88% Error = 78.12%

	Accuracy = 21.88% **Insertions: 0 Deletions: 30 Substitutions: 45**
sentence37	Words: 104 Correct: 19 Errors: 85 Percent correct = 18.27% Error = 81.73% Accuracy = 18.27% **Insertions: 0 Deletions: 24 Substitutions: 61**
sentence38	Words: 61 Correct: 11 Errors: 55 Percent correct = 18.03% Error = 90.16% Accuracy = 9.84% **Insertions: 5 Deletions: 12 Substitutions: 38**
sentence39	Words: 77 Correct: 12 Errors: 65 Percent correct = 15.58% Error = 84.42% Accuracy = 15.58% **Insertions: 0 Deletions: 18 Substitutions: 47**
sentence40	Words: 54 Correct: 11 Errors: 44 Percent correct = 20.37% Error = 81.48% Accuracy = 18.52% **Insertions: 1 Deletions: 17 Substitutions: 26**
sentence41	Words: 145 Correct: 29 Errors: 119 Percent correct = 20.00% Error = 82.07% Accuracy = 17.93% **Insertions: 3 Deletions: 36 Substitutions: 80**
sentence42	Words: 57 Correct: 7 Errors: 50 Percent correct = 12.28% Error = 87.72% Accuracy = 12.28% **Insertions: 0 Deletions: 15 Substitutions: 35**
sentence43	Words: 121 Correct: 24 Errors: 98 Percent correct = 19.83% Error = 80.99% Accuracy = 19.01% **Insertions: 1 Deletions: 36 Substitutions: 61**
sentence44	Words: 145 Correct: 31 Errors: 114 Percent correct = 21.38% Error = 78.62% Accuracy = 21.38% **Insertions: 0 Deletions: 51 Substitutions: 63**
sentence45	Words: 88 Correct: 15 Errors: 73 Percent correct = 17.05% Error = 82.95% Accuracy = 17.05% **Insertions: 0 Deletions: 39 Substitutions: 34**
sentence46	Words: 115 Correct: 28 Errors: 87 Percent correct = 24.35% Error = 75.65% Accuracy = 24.35% **Insertions: 0 Deletions: 40 Substitutions: 47**
sentence47	Words: 127 Correct: 20 Errors: 109 Percent correct = 15.75% Error = 85.83% Accuracy = 14.17% **Insertions: 2 Deletions: 22 Substitutions: 85**
sentence48	Words: 52 Correct: 9 Errors: 43 Percent correct = 17.31% Error = 82.69% Accuracy = 17.31% **Insertions: 0 Deletions: 12 Substitutions: 31**
sentence49	Words: 61 Correct: 11 Errors: 54 Percent correct = 18.03% Error = 88.52% Accuracy = 11.48% **Insertions: 4 Deletions: 4 Substitutions: 46**
sentence50	Words: 57 Correct: 14 Errors: 46 Percent correct = 24.56% Error = 80.70% Accuracy = 19.30% **Insertions: 3 Deletions: 8 Substitutions: 35**

TOTAL Words: 4470 Correct: 894 TOTAL Percent correct = 20%

Live

sentence1	Words: 94 Correct: 24 Errors: 70 Percent correct = 25.53% Error = 74.47% Accuracy = 25.53% **Insertions: 0 Deletions: 14 Substitutions: 56**
sentence2	Words: 97 Correct: 25 Errors: 83 Percent correct = 25.77% Error = 85.57% Accuracy = 14.43% **Insertions: 11 Deletions: 8 Substitutions: 64**

sentence3	Words: 79 Correct: 27 Errors: 60 Percent correct = 34.18% Error = 75.95% Accuracy = 24.05% **Insertions: 8 Deletions: 3 Substitutions: 49**
sentence4	Words: 63 Correct: 25 Errors: 44 Percent correct = 39.68% Error = 69.84% Accuracy = 30.16% **Insertions: 6 Deletions: 8 Substitutions: 30**
sentence5	Words: 99 Correct: 22 Errors: 89 Percent correct = 22.22% Error = 89.90% Accuracy = 10.10% **Insertions: 12 Deletions: 2 Substitutions: 75**
sentence6	Words: 84 Correct: 25 Errors: 80 Percent correct = 29.76% Error = 95.24% Accuracy = 4.76% **Insertions: 21 Deletions: 0 Substitutions: 59**
sentence7	Words: 112 Correct: 32 Errors: 96 Percent correct = 28.57% Error = 85.71% Accuracy = 14.29% **Insertions: 16 Deletions: 5 Substitutions: 75**
sentence8	Words: 93 Correct: 34 Errors: 59 Percent correct = 36.56% Error = 63.44% Accuracy = 36.56% **9**
sentence9	Words: 102 Correct: 25 Errors: 81 Percent correct = 24.51% Error = 79.41% Accuracy = 20.59% **Insertions: 4 Deletions: 7 Substitutions: 70**
sentence10	Words: 60 Correct: 18 Errors: 47 Percent correct = 30.00% Error = 78.33% Accuracy = 21.67% **Insertions: 5 Deletions: 3 Substitutions: 39**
sentence11	Words: 93 Correct: 26 Errors: 71 Percent correct = 27.96% Error = 76.34% Accuracy = 23.66% **Insertions: 4 Deletions: 12 Substitutions: 55**
sentence12	Words: 100 Correct: 28 Errors: 82 Percent correct = 28.00% Error = 82.00% Accuracy = 18.00% **Insertions: 10 Deletions: 9 Substitutions: 63**
sentence13	Words: 54 Correct: 17 Errors: 45 Percent correct = 31.48% Error = 83.33% Accuracy = 16.67% **Insertions: 8 Deletions: 2 Substitutions: 35**
sentence14	Words: 103 Correct: 26 Errors: 87 Percent correct = 25.24% Error = 84.47% Accuracy = 15.53% **Insertions: 10 Deletions: 9 Substitutions: 68**
sentence15	Words: 92 Correct: 35 Errors: 80 Percent correct = 38.04% Error = 86.96% Accuracy = 13.04% **Insertions: 23 Deletions: 2 Substitutions: 55**
sentence16	Words: 98 Correct: 22 Errors: 78 Percent correct = 22.45% Error = 79.59% Accuracy = 20.41% **Insertions: 2 Deletions: 15 Substitutions: 61**
sentence17	Words: 75 Correct: 16 Errors: 61 Percent correct = 21.33% Error = 81.33% Accuracy = 18.67% **Insertions: 2 Deletions: 8 Substitutions: 51**
sentence18	Words: 84 Correct: 27 Errors: 61 Percent correct = 32.14% Error = 72.62% Accuracy = 27.38% **Insertions: 4 Deletions: 11 Substitutions: 46**
sentence19	Words: 110 Correct: 34 Errors: 80 Percent correct = 30.91% Error = 72.73% Accuracy = 27.27% **Insertions: 4 Deletions: 10 Substitutions: 66**
sentence20	Words: 71 Correct: 21 Errors: 51 Percent correct = 29.58% Error = 71.83% Accuracy = 28.17% **Insertions: 1 Deletions: 11 Substitutions: 39**
sentence21	Words: 111 Correct: 25 Errors: 87 Percent correct = 22.52% Error = 78.38% Accuracy = 21.62%

	Insertions: 1 Deletions: 23 Substitutions: 63
sentence22	Words: 63 Correct: 9 Errors: 55 Percent correct = 14.29% Error = 87.30% Accuracy = 12.70% **Insertions: 1 Deletions: 2 Substitutions: 52**
sentence23	Words: 65 Correct: 24 Errors: 42 Percent correct = 36.92% Error = 64.62% Accuracy = 35.38% **Insertions: 1 Deletions: 11 Substitutions: 30**
sentence24	Words: 74 Correct: 12 Errors: 65 Percent correct = 16.22% Error = 87.84% Accuracy = 12.16% **Insertions: 3 Deletions: 18 Substitutions: 44**
sentence25	Words: 97 Correct: 25 Errors: 72 Percent correct = 25.77% Error = 74.23% Accuracy = 25.77% **Insertions: 0 Deletions: 20 Substitutions: 52**
sentence26	Words: 73 Correct: 14 Errors: 60 Percent correct = 19.18% Error = 82.19% Accuracy = 17.81% **Insertions: 1 Deletions: 18 Substitutions: 41**
sentence27	Words: 90 Correct: 21 Errors: 70 Percent correct = 23.33% Error = 77.78% Accuracy = 22.22% **Insertions: 1 Deletions: 18 Substitutions: 51**
sentence28	Words: 63 Correct: 14 Errors: 51 Percent correct = 22.22% Error = 80.95% Accuracy = 19.05% **Insertions: 2 Deletions: 5 Substitutions: 44**
sentence29	Words: 67 Correct: 20 Errors: 49 Percent correct = 29.85% Error = 73.13% Accuracy = 26.87% **Insertions: 2 Deletions: 7 Substitutions: 40**
sentence30	Words: 65 Correct: 15 Errors: 52 Percent correct = 23.08% Error = 80.00% Accuracy = 20.00% **Insertions: 2 Deletions: 15 Substitutions: 35**
sentence31	Words: 47 Correct: 12 Errors: 36 Percent correct = 25.53% Error = 76.60% Accuracy = 23.40% **Insertions: 1 Deletions: 2 Substitutions: 33**

TOTAL Words: 2578 Correct: 700 Errors: 2044
TOTAL Percent correct = 27.15% Error = 79.29% Accuracy = 20.71%
TOTAL Insertions: 166 Deletions: 289 Substitutions: 1589

Transcription of Sample audio:

दुनिया के देशो में भारत का उच्च स्थान वैज्ञानिक व प्रोद्योगिक उत्कृष्टता के जरिये सामाजिक आर्थिक परिवर्तन लाने की हमारी क्षमता पर निर्भर करता है

*** D u n I J A K E D ʃE O M E Bʰ A rə T k a U ʃT I T ʰ a N V ɛ g J a n I K ʊ A P R ɔ D J O g i K U T K ʃ T a k E ə R I E s a M A U K A T ʰ I k P ə R I ʊ ə R T ə N L a N E K I H M a R i K ə M T a P ə ɾ N I r B ə r K ə ɾ t A H ɛ

SIL ʃ u n *** *** *** *** *** *** *** ə Kʰ N Õ SIL H r *** R k a *** *** *** K ɾ K a D SIL SIL I a n F I ə N Ã ə ɾ S ə H i ** G ʰ ʃ T ʃ ə ʃ J a k *** *** *** I Ĺ s a *** *** *** *** S ə r *** *** k *** *** *** A ə I A ə *** a *** *** *** O J A a *** i ***ʃT õ K a *** *** *** *** S r F I r A ɾ t *** K ə

Percent correct = 22.88%

मेरी सरकार का सतत प्रयास रहेगा कि अनुसंधान और विकास पर खर्च जीडीपी की एक एक फीसदी से बढ़कर दो फीसदी हो जाये

M E R I S ə R k A R k a ə T ə T P r ə J a s R ə H E g a K I A N U õSN Dʰ a N ɔ R ʊ i K A s ə R Kʰ ə ɾ TʃDʒ I d i P I K i E E K f I S E d I S ə ɪB k ə ɾ D O F I S D I H O ʒDA E

*** *** SIL K Dʰ E k ə ɾ k a *** *** *** R R R r KʃR a s *** ə I A L a *** R K I D

DʒD J TʃaI *** *** *** G iI *** *** s ɹə T R AI ɾ *** *** *** *** iI *** E ɾ iI D T □f Bʰ
L I d *** *** *** *** ə H k ə *** □K E K □P J □T P U J
Percent correct = 21.98%

ग्यारवी योजना अवधि के दौरान अनुसंधान और विकास में सार्वजनिक निवेश प्रति वर्ष बीस से पच्चीसफीसदी की दर से बढ़ा है
*** g j aI R hə ʊ lI J O ɡn A aɔb Dʰ lI k E dɔ R A n A N UI SN Dʰ AI NɔR ʊ lI K aI S M E S aI ʊrDʒ ə N lI K N iɔb e ʃ P R ə T lI ʊ ə ɾ ʃ B lI S S e Pə TʃTʃ lI S F lI S D lI K lI ɑɹ S E B ə ʈ aI H ɛ
SIL H j aI □h *** *** *** *** *** R n *** aI *** *** *** k ə d Bʰ ʈ lI n *** AI J AI I SIL A S D ɾ ʊ *** *** aI *** *** D H aI r L AI D RəJD iI □e *** *** *** *** *** K UI ə *** ʃ *** *** R H e *** KʃI L H E E S E □F E R d ə *** *** *** A ɾ J aI *** ***
Percent correct = 21.28%

सरकार ने इनोवेशन इन सैन पर्स्युट और इंस्पायर रिसर्च अर्थात इंस्पायरस की सफलतापूर्वक शुरू की है और इसके तहत अब तक विज्ञान विषय के पांच लाख से अधिक छात्रों को पुरस्कार दिया जा चुका है
*** *** S ə ɾ k aI r N E lI n Oɔb E ʃ ə N lI N S □N P ə ɾ s j UI ʈ ɔ R iI N S p AI ə R ɖ r lI sə R TʃA R Tʰ AI T lI N S p aI əJɾ lI s K lI M sə F ə L T AI p UI ɒ ə KʃUI r UI K lI ɑhɔ R iI S K E Tɛ H ə T A B T ə k ʊ lI g J AI nʊ iI ʃ ə j *** *** K E P AI ʄl aI *** kˢ E A Dʰ iI K Jʰ AI t R O k O P U RəS K aI r D lI J AIʒ AI TʃUI K A Hɛ

SIL T K ə ɾ k aI r *** ***ə n *** *** *** *** S K UI R R J □ə TʃAI s j *** O D I iI *** □p *** *** *** *** □r ʃ s ə T ʃ R A ə S ə D H □p aI *** *** A□s *** *** ɛ s *** *** ə *** *** ɾ p AI r AIr T□ Ãr *** *** *** h *** *** *** iI ***ʊ T ʈ S B ə lI E D̂ AI Kk *** *** ə ʈ ɾ n F i□E l j R D Ĩ□UI E ʄl aI T SILs K ə H iI *** *** *** t *** *** k *** AI□B D SIL R aI r *** *** SIlɔ Bʰ ə U E ʃ lI B ə
Percent correct = 24.50%

विज्ञान और प्रोद्योगिकी क्षेत्र में नवाचार इनोवेशन को सुसाध्य करने के लिए कई नए संस्थागत प्रयास किये गए है
*** ʊ iI g j AI NɔR P r ɔ D J O g iI K iI ʃKE T R ə M e n ə ʊ aI ʃ aI R lI N Oɔ E ʃ ə N K O S UI S aI ʰDl ə k ə ɾ N E K E L iI E Kə lI Nə lI SɜN S Tʰ A g ə T P R ə j aI S K iI Eg ə J E H ɛ

SIL H iI T j *** *** *** *** *** r *** KʃI ə H iI *** iI *** S lI ɑKɾ e n *** *** aI B aI O DZ AI Hʊ *** I TʃI □ʃ D Ãʈ K aI Fɖ R k ə *** *** *** R H Ĺ iI D PT *** *** *** *** *** *** *** *** *** ɾ AI AI J AI j ɑàU iI *** *** *** *** *** *** ***
Percent correct = 21.35%

Transcription of Live audio:

भारत के पश्चिमी तट पर इस्थित मुम्बई भारतीय राज्य महाराष्ट्र की राजधानी है *** *** *** *** Bʰ a ɾ ə t k E P ə ʃ tʃ I m l Tə ʈ P ə *** R i S Tʰ i T m u m b ə l Bʰ a ɾ t l ɻ R a ɡ J ə m ə H a R a ʃ ʈ ɾ ə K i R a ɡ Dʰ a *** N l Hɛ SIL S ʈ M P a Rə t k ə ɾ ə *** tʃ *** m E I I ɾ ə ə P B i I ɾ i *** m u m b ə O P a *** t U E L a *** *** m *** O a L a *** T SʈL H i L a U K a OʒD J SIL Percent correct = 39.68%
भारत के राष्ट्रपति राश प्रमुख और भारत के प्रथम नागरिक हैं साथ ही भारतीय सशत्र सेनाओं के सेनापति भी हैं Bʰ ə ɾ ə t k E r a ʃ ʈ ɾ ə P ə T l R a ʃ P ɾ ə m U Kʰ ɔ R Bʰ a ɾ ə t k E P ɾ ə Tʰ ə M n a R I K H S A Tʰ H i Bʰ a ɾ t l J Sə ʃ ə ʃ t ɾ ə S e n a Õ k E P ɾ ə M U Kʰ S E n a Pə t i Bʰ l H *** SIL D A t k ɡ r a ʃ ʈ ɾ ə *** T ɖ I L a ʃ ʈ S ə m *** *** M Õ P a PA t k *** *** ə Dʰ O N A n a ʰDʒ l Dʰ ə ʃ ʈ P A T i P a ə t ə K ɾ ə S ə *** t *** ə K e n a ʊ k *** *** ɾ P U I O G Dʒ n a ɾ ə t i *** ʊ SIL ʈ Percent correct = 36.56%
आई आई टी कानपुर मुख्य रूप से विज्ञान एवं अभियांत्रिकी में सौ तथा स्नातक शिक्षा पर केन्द्रित एक प्रमुख भारतीय तकनीकी संस्थान बनकर उभरा है *** a l a ʈ i k a n p U R m U ʰKʃ ə R u P S E ʊ i g g J a n *** E ʊ ə m a Bʰ I J a n T ɾ i k Im *** *** Ẽ S o T ə Tʰ A I S n A Tə K S I Tʃʰ A Pə R K EN d ɾ l t E K p ɾ ə M U Kʰ Bʰ A ɾ t i J Tə K N i K l Sə̃S Tʰ A N Bə N k E U Bʰ R A Hɛ SIL a l a ʈ l k a n pə M m UK j *** Õ u *** ʃ I ɾ i *** Dʰ K a nə Tʃ M ə m a B ɖ l a n ***ʰD k Dʒ m G ɛ ʈ Bʰ o ɾ M ʈ B ə ʃ n O ɾ ə ʃ I K ɾ ə ɾ ə K E N d P E t ɾ I p *** *** U Bʰ O P M ə t i I ɾ ə O ʊ i ***ə ɾ H ə M ÃK ÕO S k *** *** *** A ɔ ə SIL Percent correct = 30.91%
सेना पदक भारतीय सेना के सभी श्रेणी के सदस्यों को सम्मानित करने के लिए दिया जाता है *** s e n A Pə D ə K bʰ a ɾ t I J ə S e n a k E sə Bʰ l ʃ ɾ ə ɳ i K E Sə D ə S J Õ K O S ə M m A N i T Kə R N E K E L I j E D I J a ɡ A T A Hɛ SIL s e n ʈ ɾ ə L ə *** bʰ a *** t T U ə K e n a k I s *** *** *** ʃ I TʃE i B I I ʈ ɾ ə *** K ʃ A U A U O m ***ʃ i ʈ ɾ ə *** L l ʈ l B l j *** *** ***ə a O ʈ A ʈ J SIL Percent correct = 29.58%
राष्ट्रीय स्वास्थ्य बीमा योजना भारती गरीब के लिए एक सरकारी स्वास्थ्य बीमा योजना है *** r a ʃ ʈ ɾ l j ə s ʊ a S ʰt ə B l m A J O ʒDn A ʰBa R ə t l g ə R i B K E L i J E E K S ə ɾ k a R i Sʊ A S ʰt B i m A J o ʒDN a H ɛ SIL r a ʃ *** *** F j ə s ʊ a *** P U Ɛɻ m *** l U ʃ n ə M a ***ə t ɖ I ə *** i *** ʈ l G i *** *** ʃ ʈ ɾ ə P k a *** i P U ə Tʰ ʊ F i m l ʰGo *** ʃ a ʈ SIL Percent correct = 36.92%

3

Thapar University Patiala

Project Progress Report of Thapar University Patiala	
A. General	
A.1 Name of the Project	**Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages**
Reference Letter No.	11(6)/2011-HCC(TDIL)dated 23-12-2011
A.2 Executing Agency	Thapar University
A.3 Chief Investigator with Designation	R. K. Sharma Professor, School of Mathematics and ComputerApplications, Thapar University Patiala (Punjab) 147 004
A.4 Project staff with Qualification (engaged at different periods of time during the project period)	Rupinderdeep Kaur, M.E. (S.E.), Ph.D. student; Baljinder Badhan, B.E. (Computer Science): Project Associate; Amandeep Kaur, Ph.D. : Project Associate (Part time); Simerjeet Kaur, Ph.D. : Project Associate (Part time); Kamal Meet Kaur, Ph.D. : Project Associate (Part time); Virender Kadyan, M.Tech. : Project Associate (Part time) Sakshi Mittal, M.Tech. (CSA) : M.Tech. student Minakshi Bansal : M.Tech. student Neeshu Agarwal : M.Tech. student
A.5 Total Cost of the Project as approved by DIT	
i) Original	30.015 Lakh
ii) Revised, if any	None
A.6 Date of starting (Indicate date of first sanction)	23-12-2011
A.7 Date of Completion	
i) Original	22-12-2013

ii) Revised, if any	31-03-2015
A.8 Date on which last progress report was submitted	28-02-2013

B. Technical	
B.1 Work Progress (Details are given in Technical Report at Appendix - I)	• Database has been collected in three different modes as per the following details: i) Read mode speech: 15 hours and 44 minutes ii) Lecture mode speech: 5 hours and 8minutes iii) Conversational mode speech: 4 hours and 19 minutes • Transcription of this database has been carried out using IPA chart: i) Read mode speech: 5 hours and 5minutes ii) Lecture mode Speech: 2 hours and 51minutes iii) Conversational mode speech: 2 hours and 33 minutes • Extraction of prosody knowledge • Development of phonetic engine
i) Proposed plan of work highlighting the action to be taken to achieve the originally proposed targets	We have achieved the following objectives in this project : • Speech Database Development • Manual Transcription of various modes of speech • Development of Phonetic Engine • Improvements in the performance of Phonetic Engine • Manual as well as Automated Pitch Accent Marking of read speech • Manual as well as Automated Break Index Marking of read speech • Manual as well as Automated Syllabification of read speech
C. Project Outcomes	
C.1Papers Published	We shall work on bringing out publication(s) from the project work.

C.2 Development of Database	• Read mode speech: 15 hours and 44 minutes • Lecture mode speech: 5 hours and 8 minutes • Conversational mode speech: 4 hours and 19 minutes
C.3 Tools and systems developed	The systems have been developed for implementation of Phonetic Engine and have been shared with IIT Guwahati.

Appendix 3.1

Detailed Technical Report of Thapar University Patiala

Progress Summary Report of Thapar University Patiala

3.1 Background

Speech is a natural and very easy way of exchanging information. If used as a medium to interact with the computer, it can solve various problems. For this, some speech interfaces such as speech synthesizer and speech recognizer are required. Speech recognition and speech synthesis both require phonetic transcription. In speech recognition, speech is provided as an input to system and then corresponding phonetic transcription is generated by the system as output. Phonetic Engine (PE) is such a module that uses the acoustic phonetic information present in the speech signal for converting the speech signal into symbolic form. This symbolic form is nothing but the basic sound units present in the spoken utterances of speech signal. These basic sound units can be represented in symbolic form using International Phonetic Alphabet (IPA) transcription standard. Acoustic phonetic information means that the PE will use the sounds of phones of spoken utterances and these sounds are represented in the symbolic form.

3.2 Database collection and transcription

3.2.1 Database collection: In this project, data has been collected in three different modes of speech, namely, read mode, lecture mode and conversational mode.

i) **Read Mode Speech:**For this mode of speech, data has been collected from 12 native Punjabi speakers for a duration of about 14 hours. This data has been recorded in a normal room environment using amicrophone channel maintained at a sampling frequency of 22050 Hz. We also collected around 1 hour of data from a radiochannel, *Sabrang* Radio. The data collected from this channel has a sampling frequency of 48 KHz and a bit rate of16 bits per sample.

Total time of collected data: 15 hours and 44 minutes.

ii) **Lecture Mode:** The data for the lecture mode has been taken from the radio channel,*Punjabi Radio USA* and also from *YouTube*. This data is available in public domain. The recording of this data has been done with a sampling frequencyof 48 KHz and a bit rate of 16 bits per sample.

Total time of collected data: 5 hours and 44 minutes.

iii) **Conversational Mode:**For this mode of speech, data has been recorded with the help of native Punjabi speakers.The duration of the conversation between native Punjabi speakers is approximately 4 hours. Out of this 4 hours data, 2 hours of data has been recorded in a normal room environment using a microphone channel maintained at a sampling frequency of 48KHz. The remaining 2 hours of data has been recorded in an open environment using a microphone channel maintained at a sampling frequency of 48KHz. Besides this, approximately 30 minutesof data has been collected in this mode from a Punjabi news channel. The data from news channel has a sampling frequency of 48 KHz and a bit rate of 16 bits per sample.

Total time of collected data: 4 hours and 19 minutes.

3.2.2 DatabaseTranscription:

The data that has been collected is transcribed using the International Phonetic Alphabet (IPA) chart. The transcription of 5 hours and 5 minutes of read speech, 2 hours and 32 minutes of lecture speech, and 2 hours and 29 minutes of conversational speech has been completed.

The database, which has been shared with IIIT-Hyderabad is divided into 3 parts, namely, SPC-1, SPC-2, and SPC-3. The database SPC-1 contains the .wav files of read mode speech, lecture mode speech, and conversational mode speech for vocabulary purpose. The database SPC-2 contains all data collected, of each kind of speech mode and the database SPC-3 contains the transcribed data of all 3 kinds of speech modes. In SPC-3, each wave file also has its corresponding transcription '.ph' file.

3.3 Development of phonetic engine

3.3.1 Transcription

Transcription of read speech, lecture speech, and conversational speech has been done using International Phonetic Alphabet (IPA) chart, which is available at*http://westonruter.github.io/ipa-chart/keyboard/*. There are 36 IPA symbols including Vowels, Semi vowels,and Consonants. Apart from this, diacritics, tone and word accents, and suprasegmentals were also used in transcription. Consonants include Steps, Velar, Affricates, Nasals, Laterals, and Fricatives.

Figure 3.1 contains the transcription of a lecture mode speech. After selecting a segment, its transcription is noted down in the transcription pane using IPA chart.

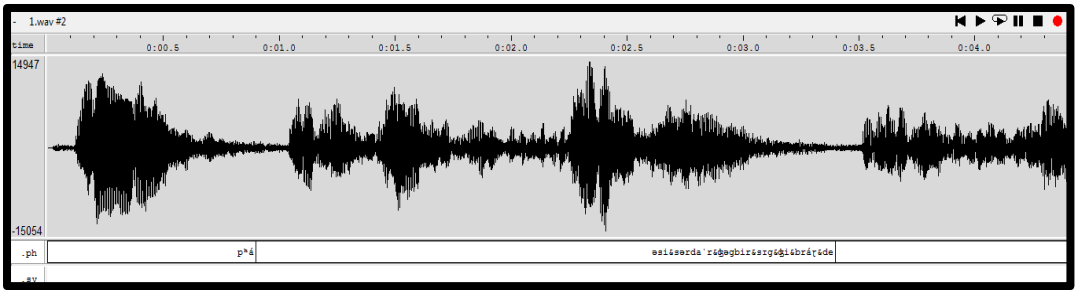


Figure 3.1: Transcription of a Lecture Speech File

3.3.2 Break Index Marking

This work is done with an objective of semi-automating the break index marking and silence removal. For doing this, each wave file is divided into overlapping frames, then energy level of each frame is computed, if it is less than 1.2 dB then it is detected as silence. Detected silence is then removed from the wave file. Figure 3. 2 contains the snapshot of a file containing time stamping of the break indices and Figure 3. 3 contains these markings on a wave file.

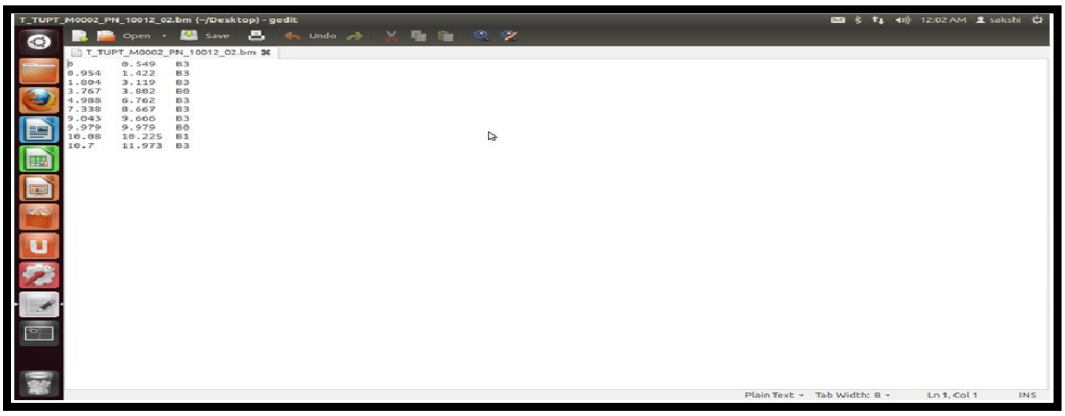


Figure 3.2: System generated break index markings

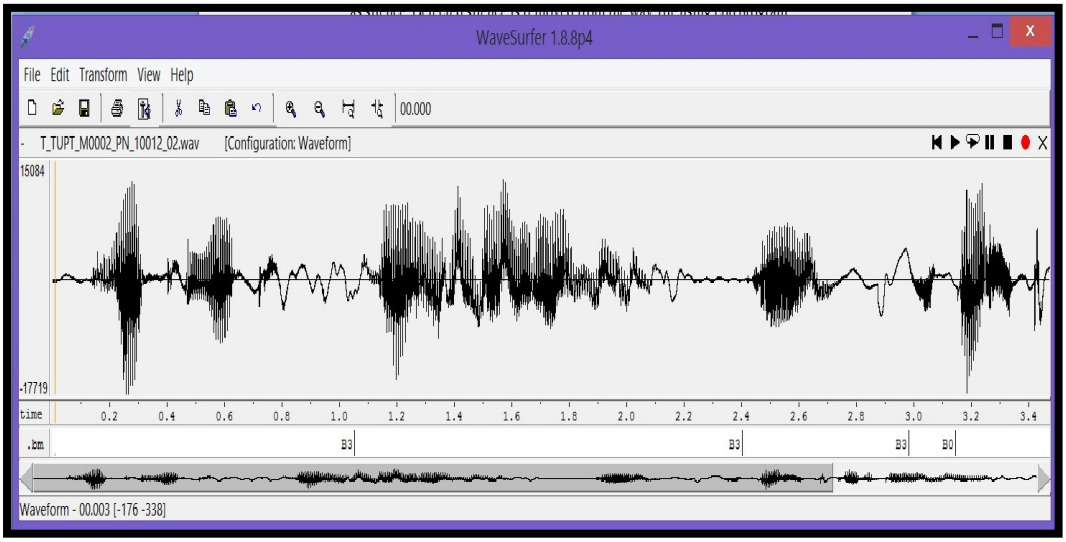


Figure 3.3:System generated break index markings corresponding to wave signal

3.3.3 Pitch Accent Marking

This work has been carried out with the objective of semi-automating the process of pitch marking. Firstly, in the whole speech, voiced and unvoiced regions are detected so that only the voiced regions are pitch marked. In a particular voiced segment of speech, pitch accent may have 7 different marks, namely, low to high (LH), high to low (HL), flat(F), *i.e.*, no change in pitch, very low to high (VLH), very high to low (VHL), low to very high (LVH) and high to very low (HVL). Zero frequency filtering technique is used to segment the speech into voiced and unvoiced region. Then, pitch marking is done for each voiced region. For pitch marking, sampling rate of the signal should not be more than 8000 Hz. If so, it is re-sampled to 8000 Hz. Finally, line fitting is done using linear regression to detect pitch variation. Figures 3.4 and 3.5 contains the results of efforts in pitch marking automation.

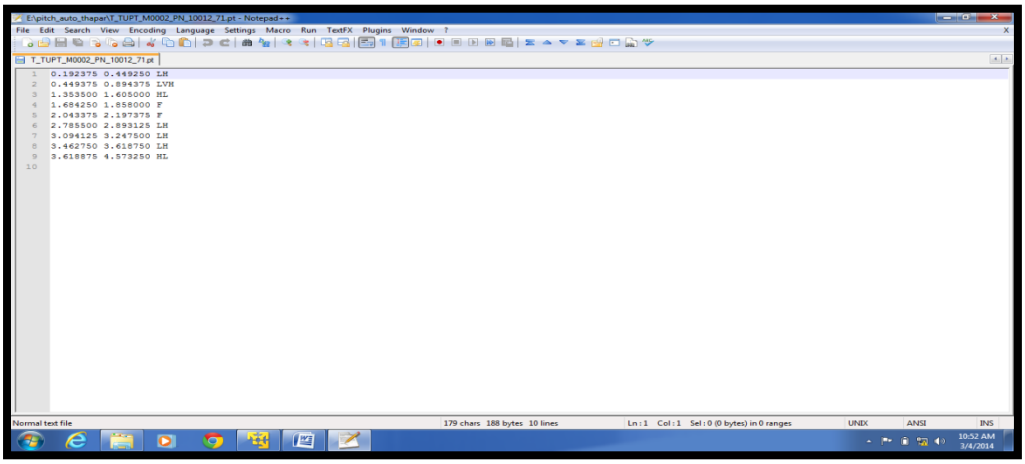


Figure 3.4: System generated pitch markings

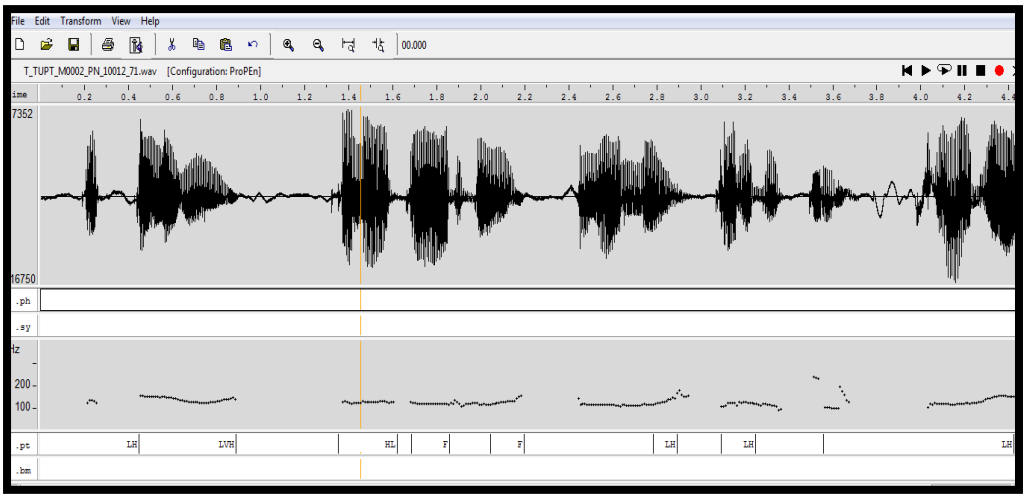


Figure 3.5:System generated pitch markings corresponding to wave signal

Illustration of pitch accent marking using a file of 4 seconds duration

Here, a .wav file of 4 seconds duration of read mode speech has been considered for pitch marking. The pitch marking has been done manually as well as generated from implemented system. After pitch marking, results of both, manual efforts and automated efforts, are depicted in Figures 3.6 and 3.7.

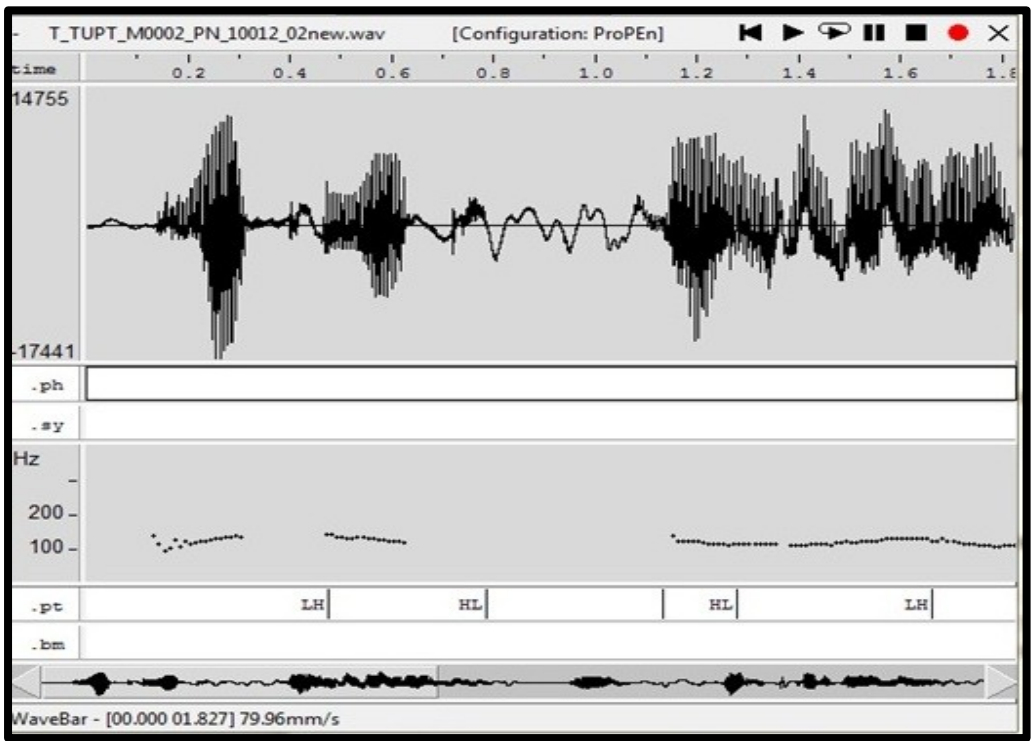


Figure 3.6: Manual pitch markings corresponding to wave signal

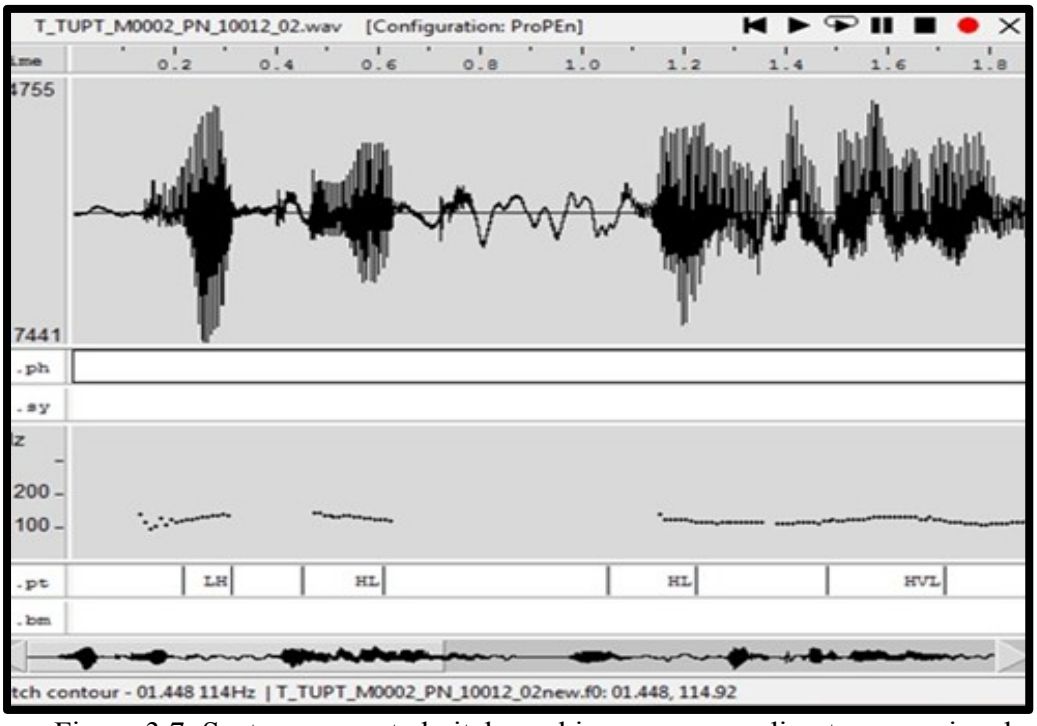


Figure 3.7: System generated pitch markings corresponding to wave signal

Results: Table 3.1 contains the results of this experimentation on the file. It has been noted that in the duration of 4 Sec., LH segment appeared 4 times if we do the marking manually. Where as in the case of automation, this segment appears only once, i.e., it has an error of 3 occurrences. Similarly the occurrences of other markings have been noted and reported in Table 3.1.

Table 3.1: Results for manual vs. system generated pitch marking

	LH	HL	F	VLH	VHL	LVH	HVL
Manual	4	3	1	0	1	0	0
Automated	1	5	0	0	1	0	1
Error	3	2	1	0	0	0	1

Illustration of pitch accent marking using a file of 15 seconds duration

Figures 3.8 and 3.9 represent the manual pitch marking and system generated pitch marking, respectively, for a file with 15 second duration. Table 3.2 depicts the error analysis of manual and system generated pitch accent marking.

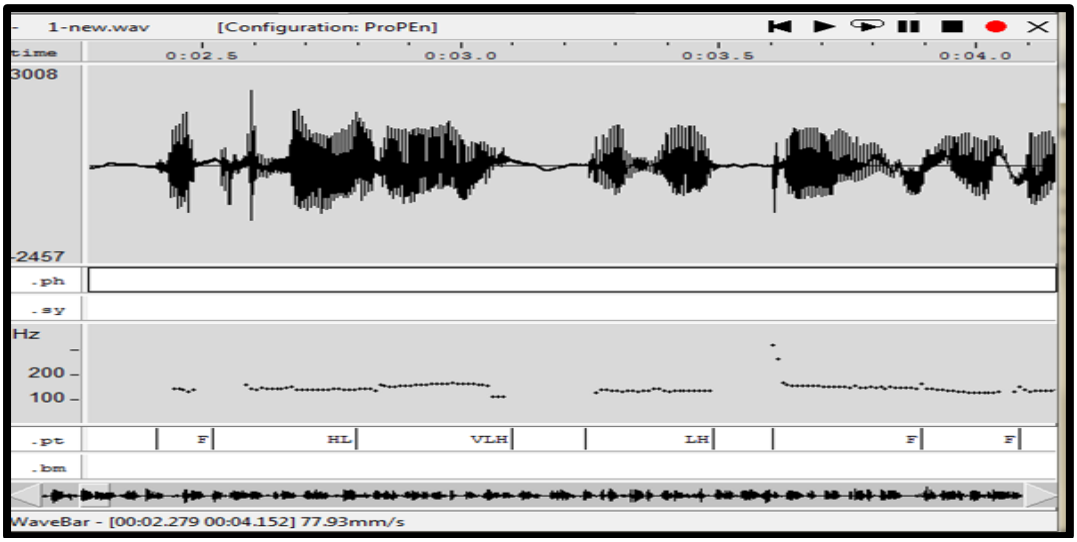


Figure 3.8: Manual Pitch markings corresponding to wave signal

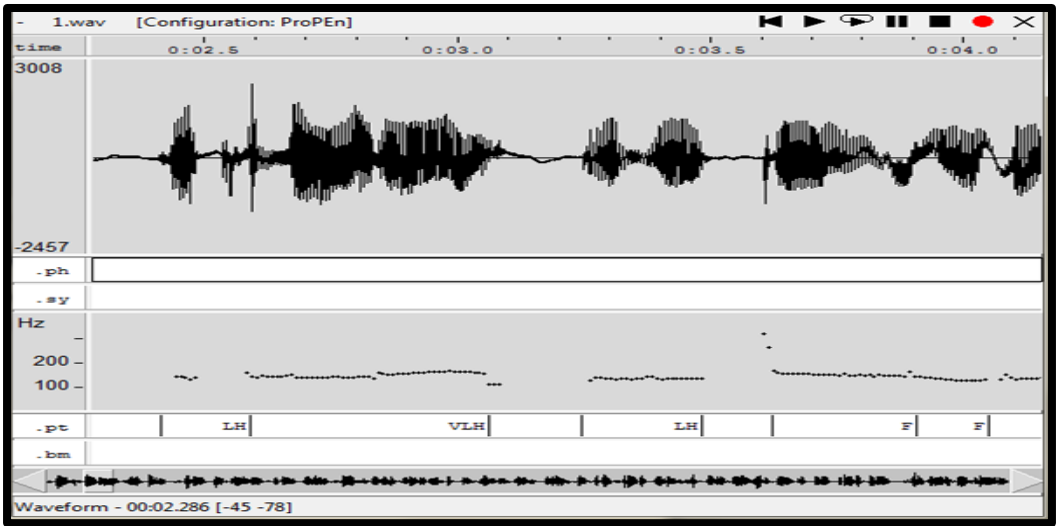


Figure 3.9: System generated Pitch Markings corresponding to wave signal

Table 3.2: Results of manual vs. system generated pitch marking

	LH	HL	F	VLH	VHL	LVH	HVL
Manual	9	12	9	1	0	0	2
Automated	7	8	8	2	1	1	2
Error	2	4	1	1	1	1	0

3.3.4 Phonetic Transcription Automation

This work has been carried out with the objective of automation of transcription. We have used the HTK toolkit in this process. Phonetic Engine (PE) is trained and tested for read speech mode, lecture speech mode, and conversational speech mode. For each mode, available transcribed data is divided into two parts. First part consists of 75% of available transcribed data, which is used for training and rest, 25% of available transcribed data is used for testing.

3.3.4.1 Phonetic Transcription Automation for different speech modes

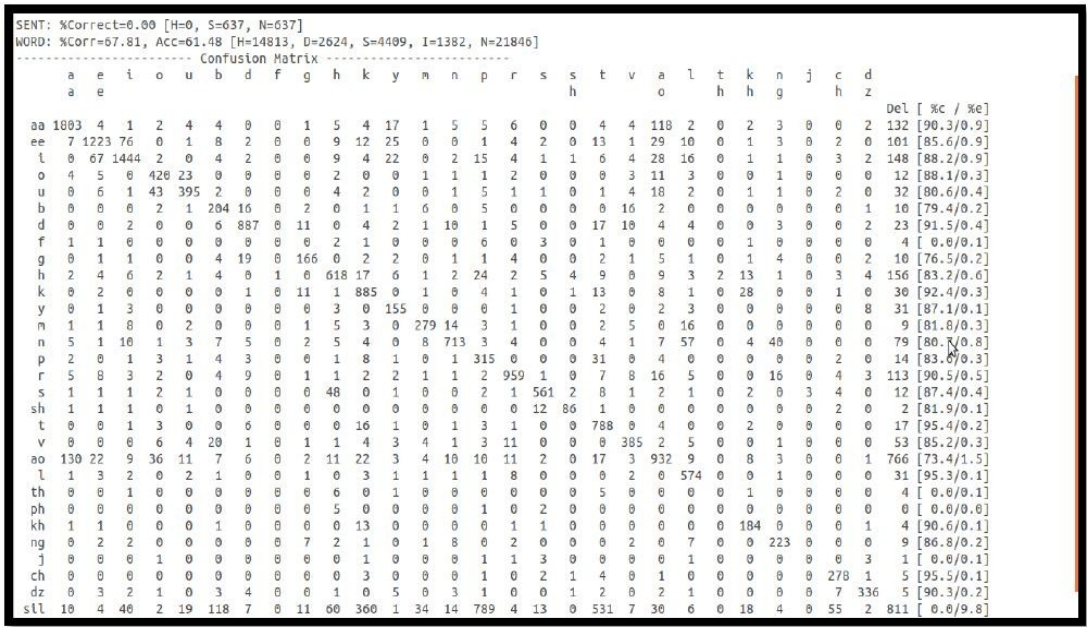
PE is trained for all the modes of speech: Read mode speech, Lecture mode speech and Conversational mode speech, as mentioned earlier. The total duration of each kind of data considered for PE automation is given in Table 3.3.

Table 3.3: Time duration of each mode of Speech

Mode	Total Duration (in Minutes)
Read	185.93
Lecture	20.05
Conversational	20.22

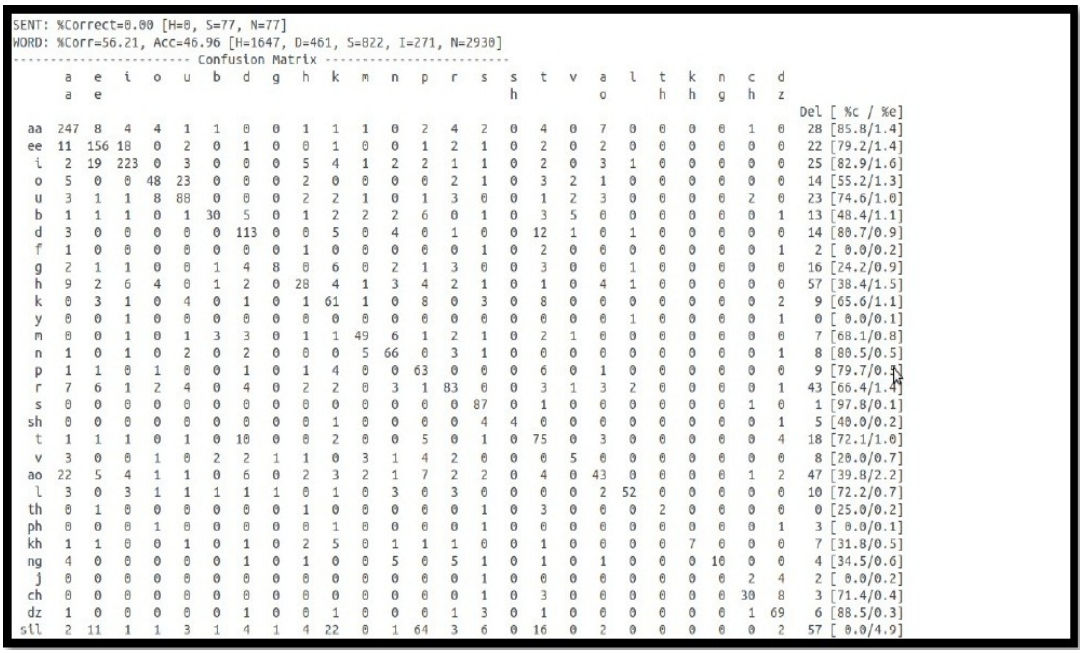
The PE is trained and tested for all three modes of speech. The results are given in the form of confusion matrix and the accuracy achieved. Confusion matrix for Read Speech mode is shown in Figure 3.10. It depicts that percentage of phonemes correctly recognized is 67.8%. Figure 3.11 shows the confusion matrix

for Lecture mode speech. This depicts that percentage of phonemes recognized correctly is 56.2%. Confusion matrix as shown in Figure 3.12 depicts that the percentage of correctly recognized phonemes is 36.7% for conversational mode speech.



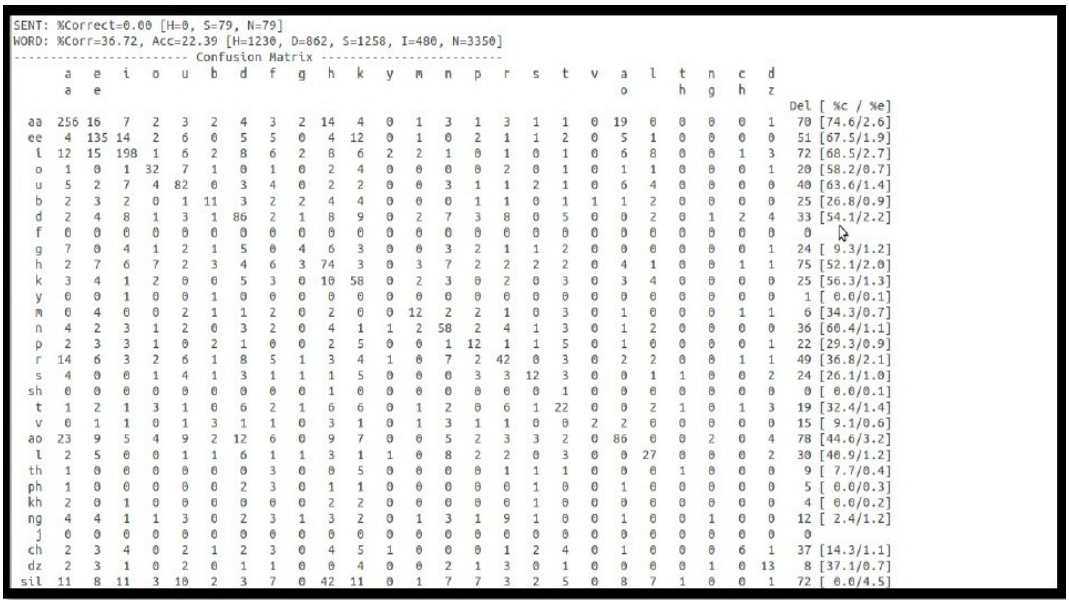
```
SENT: %Correct=0.00 [H=0, S=637, N=637]
WORD: %Corr=67.81, Acc=61.48 [H=14813, D=2624, S=4409, I=1382, N=21846]
------------------------ Confusion Matrix -------------------------
      a   e   i   o   u   b   d   f   g   h   k   y   m   n   p   r   s   s   t   v   a   l   t   k   n   j   c   d
      a   e                                                       h           o       h   h   g       h   z
                                                                                                              Del [ %c / %e]
 aa 1803  4   1   2   4   4   0   0   1   5   4  17   1   5   5   6   0   0   4   4 118   2   0   2   3   0   0   2 132 [90.3/0.9]
 ee    7 1223 76   0   1   8   2   0   0   9  12  25   0   0   1   4   2   0  13   1  29  10   0   1   3   0   2   0 101 [85.6/0.9]
  i    0  67 1444  2   0   4   2   0   0   9   4  22   0   2  15   4   1   1   6   4  28  16   0   1   1   0   3   2 148 [88.2/0.9]
  o    4   5   0 420  23   0   0   0   0   2   0   0   1   1   1   2   0   0   0   3  11   3   0   0   1   0   0   0  12 [88.1/0.3]
  u    0   6   1  43 395   2   0   0   0   4   2   0   0   1   5   1   1   0   1   4  18   2   0   1   1   0   2   0  32 [80.6/0.4]
  b    0   0   0   2   1 204  16   0   2   0   1   1   6   0   5   0   0   0   0  16   2   0   0   0   0   0   0   1  10 [79.4/0.2]
  d    0   0   2   0   0   6 887   0  11   0   4   2   1  10   1   5   0   0  17  10   4   4   0   0   3   0   0   2  23 [91.5/0.4]
  f    1   1   0   0   0   0   0   0   0   2   1   0   0   0   6   0   3   0   1   0   0   0   0   1   0   0   0   0   4 [ 0.0/0.1]
  g    0   1   1   0   0   4  19   0 166   0   2   2   0   1   1   4   0   0   2   1   5   1   0   1   4   0   0   2  10 [76.5/0.2]
  h    2   4   6   2   1   4   0   1   0 618  17   6   1   2  24   2   5   4   9   0   9   3   2  13   1   0   3   4 156 [83.2/0.6]
  k    0   2   0   0   0   0   1   0  11   1 885   0   1   0   4   1   0   1  13   0   8   1   0  28   0   0   1   0  30 [92.4/0.3]
  y    0   1   3   0   0   0   0   0   0   3   0 155   0   0   0   1   0   0   2   0   2   3   0   0   0   0   0   8  31 [87.1/0.1]
  m    1   1   8   0   2   0   0   0   1   5   3   0 279  14   3   1   0   0   2   5   0  16   0   0   0   0   0   0   9 [81.8/0.3]
  n    5   1  10   1   3   7   5   0   2   5   4   0   8 713   3   4   0   0   4   1   7  57   0   4  40   0   0   0  79 [80.7/0.8]
  p    2   0   1   3   1   4   3   0   0   1   8   1   0   1 315   0   0   0  31   0   4   0   0   0   0   0   2   0  14 [83.6/0.3]
  r    5   8   3   2   0   4   9   0   1   1   2   2   1   1   2 959   1   0   7   8  16   5   0   0  16   0   4   3 113 [90.5/0.5]
  s    1   1   1   2   1   0   0   0   0  48   0   1   0   0   2   1 561   2   8   1   2   1   0   2   0   3   4   0  12 [87.4/0.4]
 sh    1   1   1   0   1   0   0   0   0   0   0   0   0   0   0   0  12  86   1   0   0   0   0   0   0   0   2   0   2 [81.9/0.1]
  t    0   0   1   3   0   0   6   0   0   0  16   1   0   1   3   1   0   0 788   0   4   0   0   2   0   0   0   0  17 [95.4/0.2]
  v    0   0   0   6   4  20   1   0   1   1   4   3   4   1   3  11   0   0   0 385   2   5   0   0   1   0   0   0  53 [85.2/0.3]
 ao  130  22   9  36  11   7   6   0   2  11  22   3   4  10  10  11   2   0  17   3 932   9   0   8   3   0   0   1 766 [73.4/1.5]
  l    1   3   2   0   2   1   0   0   1   0   3   1   1   1   8   0   0   0   0   2   0 574   0   0   1   0   0   0  31 [95.3/0.1]
 th    0   0   1   0   0   0   0   0   0   6   0   1   0   0   0   0   0   0   5   0   0   0   0   1   0   0   0   0   4 [ 0.0/0.1]
 ph    0   0   0   0   0   0   0   0   0   5   0   0   0   0   1   0   2   0   0   0   0   0   0   0   0   0   0   0   0 [ 0.0/0.0]
 kh    1   1   0   0   0   1   0   0   0   0  13   0   0   0   0   1   1   0   0   0   0   0   0 184   0   0   0   1   4 [90.6/0.1]
 ng    0   2   2   0   0   0   0   0   7   2   1   0   1   8   0   2   0   0   0   2   0   7   0   0 223   0   0   0   9 [86.8/0.2]
  j    0   0   0   1   0   0   0   0   0   0   1   0   0   0   0   1   1   3   0   0   0   1   0   0   0   0   0   3   1 [ 0.0/0.1]
 ch    0   0   0   0   0   0   0   0   0   0   3   0   0   0   1   0   2   1   4   0   1   0   0   0   0   0 278   1   5 [95.5/0.1]
 dz    0   3   2   1   0   3   4   0   0   1   0   5   0   3   1   0   0   1   2   0   2   1   0   0   0   0   7 336   5 [90.3/0.2]
sil   18   4  46   2  19 118   7   0  11  60 360   1  34  14 789   4  13   0 531   7  30   6   0  18   4   0  55   2 811 [ 0.0/9.8]
```

Figure 3.10: Confusion matrix for read speech mode



```
SENT: %Correct=0.00 [H=0, S=77, N=77]
WORD: %Corr=56.21, Acc=46.96 [H=1647, D=461, S=822, I=271, N=2930]
------------------------ Confusion Matrix -------------------------
      a   e   i   o   u   b   d   g   h   k   m   n   p   r   s   s   t   v   a   l   t   k   n   c   d
      a   e                                           h           o       h   h   g   h   z
                                                                                              Del [ %c / %e]
 aa 247  8   4   4   1   1   0   0   1   1   1   0   2   4   2   0   4   0   7   0   0   0   0   1   0   28 [85.8/1.4]
 ee  11 156 18   0   2   0   1   0   0   1   0   0   1   2   1   0   2   0   2   0   0   0   0   0   0   22 [79.2/1.4]
  i   2  19 223  0   3   0   0   0   5   4   1   2   2   1   1   0   2   0   3   1   0   0   0   0   0   25 [82.9/1.6]
  o   5   0   0  48  23   0   0   0   2   0   0   0   0   2   1   0   3   2   1   0   0   0   0   0   0   14 [55.2/1.3]
  u   3   1   1   8  88   0   0   0   2   2   1   0   1   3   0   0   1   2   3   0   0   0   0   2   0   23 [74.6/1.0]
  b   1   1   1   0   1  30   5   0   1   2   2   2   6   0   1   0   3   5   0   0   0   0   0   0   1   13 [48.4/1.1]
  d   3   0   0   0   0   0 113   0   0   5   0   4   0   1   0   0  12   1   0   1   0   0   0   0   0   14 [80.7/0.9]
  f   1   0   0   0   0   0   0   0   1   0   0   0   0   0   1   0   2   0   0   0   0   0   0   0   1    2 [ 0.0/0.2]
  g   2   1   1   0   0   1   4   8   0   6   0   2   1   3   0   0   3   0   0   1   0   0   0   0   0   16 [24.2/0.9]
  h   9   2   6   4   0   1   2   0  28   4   1   3   4   2   1   0   1   0   4   1   0   0   0   0   0   57 [38.4/1.5]
  k   0   3   1   0   4   0   1   0   1  61   1   0   8   0   3   0   8   0   0   0   0   0   0   0   2    9 [65.6/1.1]
  y   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   1    0 [ 0.0/0.1]
  m   0   0   1   0   1   3   3   0   1   1  49   6   1   2   1   0   2   1   0   0   0   0   0   0   0    7 [68.1/0.8]
  n   1   0   1   0   2   0   2   0   0   0   5  66   0   3   1   0   0   0   0   0   0   0   0   0   1    8 [80.5/0.5]
  p   1   1   0   1   0   0   1   0   1   4   0   0  63   0   0   0   6   0   1   0   0   0   0   0   0    9 [79.7/0.8]
  r   7   6   1   2   4   0   4   0   2   2   0   3   1  83   0   0   3   1   3   2   0   0   0   0   1   43 [66.4/1.4]
  s   0   0   0   0   0   0   0   0   0   0   0   0   0   0  87   0   1   0   0   0   0   0   0   1   0    1 [97.8/0.1]
 sh   0   0   0   0   0   0   0   0   0   1   0   0   0   0   4   4   0   0   0   0   0   0   0   0   1    5 [40.0/0.2]
  t   1   1   1   0   1   0  10   0   0   2   0   0   5   0   1   0  75   0   3   0   0   0   0   0   4   18 [72.1/1.0]
  v   3   0   0   1   0   2   2   1   1   0   3   1   4   2   0   0   0   5   0   0   0   0   0   0   0    8 [20.0/0.7]
 ao  22   5   4   1   1   0   6   0   2   3   2   1   7   2   2   0   4   0  43   0   0   0   0   1   2   47 [39.8/2.2]
  l   3   0   3   1   1   1   1   1   0   1   0   3   0   0   0   0   0   0   2  52   0   0   0   0   0   10 [72.2/0.7]
 th   0   1   0   0   0   0   0   0   0   1   0   0   0   0   0   1   0   3   0   0   0   2   0   0   0    0 [25.0/0.2]
 ph   0   0   0   1   0   0   0   0   0   1   0   0   0   0   0   1   0   0   0   0   0   0   0   0   1    3 [ 0.0/0.1]
 kh   1   1   0   0   1   0   1   0   2   5   0   1   1   1   0   0   1   0   0   0   0   7   0   0   0    7 [31.8/0.5]
 ng   4   0   0   0   0   0   1   0   1   0   0   5   0   5   1   0   1   0   1   0   0   0  10   0   0    4 [34.5/0.6]
  j   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   2   4    2 [ 0.0/0.2]
 ch   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   3   0   0   0   0   0   0  30   8    3 [71.4/0.4]
 dz   1   0   0   0   0   0   1   0   0   1   0   0   0   1   3   0   1   0   0   0   0   0   0   1  69    6 [88.5/0.3]
sil   2  11   1   1   3   1   4   1   4  22   0   1  64   3   6   0  16   0   2   0   0   0   0   0   2   57 [ 0.0/4.9]
```

Figure 3.11: Confusion matrix for lecture mode speech



```
SENT: %Correct=0.00 [H=0, S=79, N=79]
WORD: %Corr=36.72, Acc=22.39 [H=1230, D=862, S=1258, I=480, N=3350]
------------------------ Confusion Matrix -------------------------
      a   e   i   o   u   b   d   f   g   h   k   y   m   n   p   r   s   t   v   a   l   t   n   c   d
      a   e                                                           o       h   g   h   z
                                                                                                  Del [ %c / %e]
 aa 256  16   7   2   3   2   4   3   2  14   4   0   1   3   1   3   1   1   0  19   0   0   0   0   1   70 [74.6/2.6]
 ee   4 135  14   2   6   0   5   5   0   4  12   0   1   0   2   1   1   2   0   5   1   0   0   0   0   51 [67.5/1.9]
  i  12  15 198   1   6   2   8   6   2   8   6   2   2   1   0   1   0   1   0   6   8   0   0   1   3   72 [68.5/2.7]
  o   1   0   1  32   7   1   0   1   0   2   4   0   0   0   0   2   0   1   0   1   1   0   0   0   1   20 [58.2/0.7]
  u   5   2   7   4  82   0   3   4   0   2   2   0   0   3   1   1   2   1   0   6   4   0   0   0   0   40 [63.6/1.4]
  b   2   3   2   0   1  11   3   2   2   4   4   0   0   0   1   1   0   1   1   1   2   0   0   0   0   25 [26.8/0.9]
  d   2   4   8   1   3   1  86   2   1   8   9   0   2   7   3   8   0   5   0   0   2   0   1   2   4   33 [54.1/2.2]
  f   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0    0
  g   7   0   4   1   2   1   5   0   4   6   3   0   0   3   2   1   1   2   0   0   0   0   0   0   1   24 [ 9.3/1.2]
  h   2   7   6   7   2   3   4   6   3  74   3   0   3   7   2   2   2   2   0   4   1   0   0   1   1   75 [52.1/2.0]
  k   3   4   1   2   0   0   5   3   0  10  58   0   2   3   0   2   0   3   0   3   4   0   0   0   0   25 [56.3/1.3]
  y   0   0   1   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0    1 [ 0.0/0.1]
  m   0   4   0   0   2   1   1   2   0   2   0   0  12   2   2   1   0   3   0   1   0   0   0   1   1    6 [34.3/0.7]
  n   4   2   3   1   2   0   3   2   0   4   1   1   2  58   2   4   1   3   0   1   2   0   0   0   0   36 [60.4/1.1]
  p   2   3   3   1   0   2   1   0   0   2   5   0   0   1  12   1   1   5   0   1   0   0   0   0   1   22 [29.3/0.9]
  r  14   6   3   2   6   1   8   5   1   3   4   1   0   7   2  42   0   3   0   2   2   0   0   1   1   49 [36.8/2.1]
  s   4   0   0   1   4   1   3   1   1   1   5   0   0   0   3   3  12   3   0   0   1   1   0   0   2   24 [26.1/1.0]
 sh   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0    0 [ 0.0/0.1]
  t   1   2   1   3   1   0   6   2   1   6   6   0   1   2   0   6   1  22   0   0   2   1   0   1   3   19 [32.4/1.4]
  v   0   1   1   0   1   3   1   1   0   3   1   0   1   3   1   1   0   0   2   2   0   0   0   0   0   15 [ 9.1/0.6]
 ao  23   9   5   4   9   2  12   6   0   9   7   0   0   5   2   3   3   2   0  86   0   0   2   0   4   78 [44.6/3.2]
  l   2   5   0   0   1   1   6   1   1   3   1   1   0   8   2   2   0   3   0   0  27   0   0   0   2   30 [40.9/1.2]
 th   1   0   0   0   0   0   0   3   0   0   5   0   0   0   0   1   1   1   0   0   0   1   0   0   0    9 [ 7.7/0.4]
 ph   1   0   0   0   0   0   2   3   0   1   1   0   0   0   0   0   1   0   0   1   0   0   0   0   0    5 [ 0.0/0.3]
 kh   2   0   1   0   0   0   0   0   0   2   2   0   0   0   0   0   1   0   0   0   0   0   0   0   0    4 [ 0.0/0.2]
 ng   4   4   1   1   3   0   2   3   1   3   2   0   1   3   1   9   1   0   0   1   0   0   1   0   0   12 [ 2.4/1.2]
  j   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0    0
 ch   2   3   4   0   2   1   2   3   0   4   5   1   0   0   0   1   2   4   0   1   0   0   0   6   1   37 [14.3/1.1]
 dz   2   3   1   0   2   0   1   1   0   0   4   0   0   2   1   3   0   1   0   0   0   0   1   0  13    8 [37.1/0.7]
sil  11   8  11   3  10   2   3   7   0  42  11   0   1   7   7   3   2   5   0   8   7   1   0   0   1   72 [ 0.0/4.5]
```

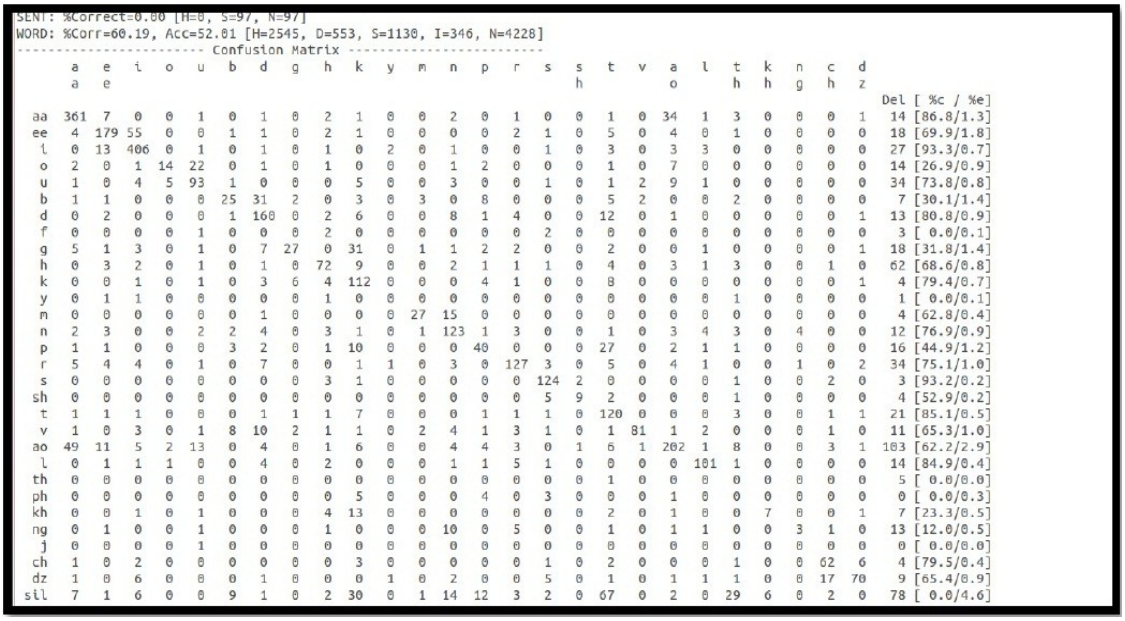
Figure 3.12: Confusion matrix for conversational mode speech

Overall accuracy of the PE is summarised in Table 3.4.

Table 3.4: Testing accuracy of PE for different speech modes

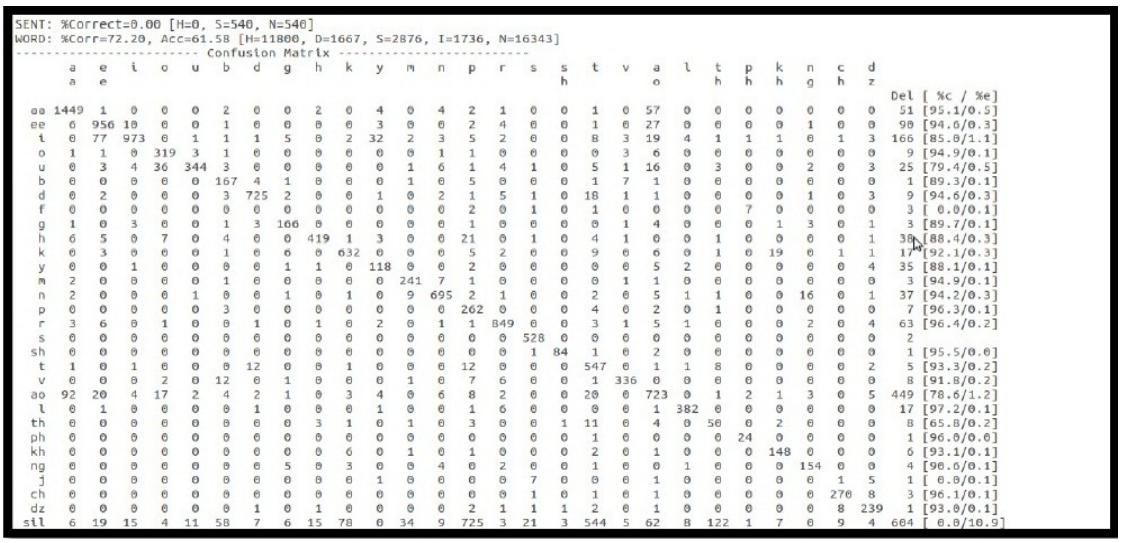
Mode of Speech	No. of Speakers	Training Data (in Minutes)	Testing Data (in Minutes)	Testing Accuracy
Read	4	138.2	47.73	61.5%
Lecture	1	15.3	4.75	47.0%
Conversational	10	15.1	5.12	22.4%

Phonetic Engine is now trained and tested for the gender, for read speech mode. For this purpose, engine was trained with 75% wave files of each category and tested for 25% wave files of the same category. The result are shown in the form of confusion matrices in Figures 3.13 and 3.14.



```
SENT: %Correct=0.00 [H=0, S=97, N=97]
WORD: %Corr=60.19, Acc=52.01 [H=2545, D=553, S=1130, I=346, N=4228]
------------------------ Confusion Matrix -------------------------
      a   e   i   o   u   b   d   g   h   k   y   m   n   p   r   s   s   t   v   a   l   t   k   n   c   d
      a   e                                                   h           o       h   h   g   h   z
                                                                                                        Del [ %c / %e]
  aa 361   7   0   0   1   0   1   0   2   1   0   0   2   0   1   0   0   1   0  34   1   3   0   0   0   1   14 [86.8/1.3]
  ee   4 179  55   0   0   1   1   0   2   1   0   0   0   0   2   1   0   5   0   4   0   1   0   0   0   0   18 [69.9/1.8]
   i   0  13 406   0   1   0   1   0   1   0   2   0   1   0   0   1   0   3   0   3   3   0   0   0   0   0   27 [93.3/0.7]
   o   2   0   1  14  22   0   1   0   1   0   0   0   1   2   0   0   0   1   0   7   0   0   0   0   0   0   14 [26.9/0.9]
   u   1   0   4   5  93   1   0   0   0   5   0   0   3   0   0   1   0   1   2   9   1   0   0   0   0   0   34 [73.8/0.8]
   b   1   1   0   0   0  25  31   2   0   3   0   3   0   8   0   0   0   5   2   0   0   2   0   0   0   0    7 [30.1/1.4]
   d   0   2   0   0   0   1 160   0   2   6   0   0   8   1   4   0   0  12   0   1   0   0   0   0   0   1   13 [80.8/0.9]
   f   0   0   0   0   1   0   0   0   2   0   0   0   0   0   0   2   0   0   0   0   0   0   0   0   0   0    3 [ 0.0/0.1]
   g   5   1   3   0   1   0   7  27   0  31   0   1   1   2   2   0   0   2   0   0   1   0   0   0   0   1   18 [31.8/1.4]
   h   0   3   2   0   1   0   1   0  72   9   0   0   2   1   1   1   0   4   0   3   1   3   0   0   1   0   62 [68.6/0.8]
   k   0   0   1   0   1   0   3   6   4 112   0   0   0   4   1   0   0   8   0   0   0   0   0   0   0   1    4 [79.4/0.7]
   y   0   1   1   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0    1 [ 0.0/0.1]
   m   0   0   0   0   0   0   1   0   0   0   0  27  15   0   0   0   0   0   0   0   0   0   0   0   0   0    4 [62.8/0.4]
   n   2   3   0   0   2   2   4   0   3   1   0   1 123   1   3   0   0   1   0   3   4   3   0   4   0   0   12 [76.9/0.9]
   p   1   1   0   0   0   3   2   0   1  10   0   0   0  40   0   0   0  27   0   2   1   1   0   0   0   0   16 [44.9/1.2]
   r   5   4   4   0   1   0   7   0   0   1   1   0   3   0 127   3   0   5   0   4   1   0   0   1   0   2   34 [75.1/1.0]
   s   0   0   0   0   0   0   0   0   3   1   0   0   0   0   0 124   2   0   0   0   0   1   0   0   2   0    3 [93.2/0.2]
  sh   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   5   9   2   0   0   0   1   0   0   0   0    4 [52.9/0.2]
   t   1   1   1   0   0   0   1   1   1   7   0   0   0   1   1   1   0 120   0   0   0   3   0   0   1   1   21 [85.1/0.5]
   v   1   0   3   0   1   8  10   2   1   1   0   2   4   1   3   1   0   1  81   1   2   0   0   0   1   0   11 [65.3/1.0]
  ao  49  11   5   2  13   0   4   0   1   6   0   0   4   4   3   0   1   6   1 202   1   8   0   0   3   1  103 [62.2/2.9]
   l   0   1   1   1   0   0   4   0   2   0   0   0   1   1   5   1   0   0   0   0 101   1   0   0   0   0   14 [84.9/0.4]
  th   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0    5 [ 0.0/0.0]
  ph   0   0   0   0   0   0   0   0   0   5   0   0   0   4   0   3   0   0   0   1   0   0   0   0   0   0    0 [ 0.0/0.3]
  kh   0   0   1   0   1   0   0   0   4  13   0   0   0   0   0   0   0   2   0   1   0   0   7   0   0   1    7 [23.3/0.5]
  ng   0   1   0   0   1   0   0   0   1   0   0   0  10   0   5   0   0   1   0   1   1   0   0   3   1   0   13 [12.0/0.5]
   j   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0    0 [ 0.0/0.0]
  ch   1   0   2   0   0   0   0   0   0   3   0   0   0   0   0   1   0   2   0   0   0   1   0   0  62   6    4 [79.5/0.4]
  dz   1   0   6   0   0   0   1   0   0   0   1   0   2   0   0   5   0   1   0   1   1   1   0   0  17  70    9 [65.4/0.9]
 sil   7   1   6   0   0   9   1   0   2  30   0   1  14  12   3   2   0  67   0   2   0  29   6   0   2   0   78 [ 0.0/4.6]
```

Figure 3.13: Confusion matrix for read speech: Female



```
SENT: %Correct=0.00 [H=0, S=540, N=540]
WORD: %Corr=72.20, Acc=61.58 [H=11800, D=1667, S=2876, I=1736, N=16343]
------------------------ Confusion Matrix -------------------------
      a   e   i   o   u   b   d   g   h   k   y   m   n   p   r   s   s   t   v   a   l   t   p   k   n   c   d
      a   e                                                   h           o       h   h   h   g   h   z
                                                                                                        Del [ %c / %e]
  aa 1449   1   0   0   0   2   0   0   2   0   4   0   4   2   1   0   0   1   0  57   0   0   0   0   0   0   0   51 [95.1/0.5]
  ee    6 956  10   0   0   1   0   0   0   0   3   0   0   2   4   0   0   1   0  27   0   0   0   0   1   0   0   98 [94.6/0.3]
   i    0  77 973   0   1   1   1   5   0   2  32   2   3   5   2   0   0   8   3  19   4   1   1   1   0   1   3  166 [85.0/1.1]
   o    1   1   0 319   3   1   0   0   0   0   0   0   1   1   0   0   0   0   3   6   0   0   0   0   0   0   0    9 [94.9/0.1]
   u    0   3   4  36 344   3   0   0   0   0   0   1   6   1   4   1   0   5   1  16   0   3   0   0   2   0   3   25 [79.4/0.5]
   b    0   0   0   0   0 167   4   1   0   0   0   0   1   0   5   0   0   0   1   7   1   0   0   0   0   0   0    1 [89.3/0.1]
   d    0   2   0   0   0   3 725   2   0   0   1   0   2   1   5   1   0  18   1   1   0   0   0   0   1   0   3    9 [94.6/0.3]
   f    0   0   0   0   0   0   0   0   0   0   0   0   0   2   0   1   0   1   0   0   0   0   7   0   0   0   0    3 [ 0.0/0.1]
   g    1   0   3   0   0   1   3 166   0   0   0   0   0   1   0   0   0   0   1   4   0   0   0   1   3   0   1    3 [89.7/0.1]
   h    6   5   0   7   0   4   0   0 419   1   3   0   0  21   0   1   0   4   1   0   0   1   0   0   0   0   1   38 [88.4/0.3]
   k    0   3   0   0   0   1   0   6   0 632   0   0   0   5   2   0   0   9   0   6   0   1   0  19   0   1   1   17 [92.1/0.3]
   y    0   0   1   0   0   0   0   1   1   0 118   0   0   2   0   0   0   0   0   5   2   0   0   0   0   0   4   35 [88.1/0.1]
   m    2   0   0   0   0   1   0   0   0   0   0 241   7   1   0   0   0   0   1   1   0   0   0   0   0   0   0    3 [94.9/0.1]
   n    2   0   0   0   1   0   0   1   0   1   0   9 695   2   1   0   0   2   0   5   1   1   0   0  16   0   1   37 [94.2/0.3]
   p    0   0   0   0   0   3   0   0   0   0   0   0   0 262   0   0   0   4   0   2   0   1   0   0   0   0   0    7 [96.3/0.1]
   r    3   6   0   1   0   0   1   0   1   0   2   0   1   1 849   0   0   3   1   5   1   0   0   0   2   0   4   63 [96.4/0.2]
   s    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 528   0   0   0   0   0   0   0   0   0   0   0    2
  sh    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1  84   1   0   2   0   0   0   0   0   0   0    1 [95.5/0.0]
   t    1   0   1   0   0   0  12   0   0   1   0   0   0  12   0   0   0 547   0   1   1   8   0   0   0   0   2    5 [93.3/0.2]
   v    0   0   0   2   0  12   0   1   0   0   0   1   0   7   6   0   0   1 336   0   0   0   0   0   0   0   0    8 [91.8/0.2]
  ao   92  20   4  17   2   4   2   1   0   3   4   0   6   8   2   0   0  20   0 723   0   1   2   1   3   0   5  449 [78.6/1.2]
   l    0   1   0   0   0   0   1   0   0   0   1   0   0   1   6   0   0   0   0   1 382   0   0   0   0   0   0   17 [97.2/0.1]
  th    0   0   0   0   0   0   0   0   0   3   1   0   1   0   3   0   0   1  11   0   4   0  50   0   2   0   0    8 [65.8/0.2]
  ph    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0  24   0   0   0    1 [96.0/0.0]
  kh    0   0   0   0   0   0   0   0   0   0   6   0   1   0   1   0   0   0   2   0   1   0   0   0 148   0   0    6 [93.1/0.1]
  ng    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   1   0   0   0 154   0    4 [98.6/0.1]
   j    0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   7   0   0   0   1   0   0   0   0   0   1    5 [ 0.0/0.1]
  ch    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   1   0   1   0   0   0   0   0 270    8 [96.1/0.1]
  dz    0   0   0   0   0   0   1   0   1   0   0   0   0   2   1   1   1   2   0   1   0   0   0   0   0   8 239    1 [93.0/0.1]
 sil    6  19  15   4  11  58   7   6  15  78   0  34   9 725   3  21   3 544   5  62   8 122   1   7   0   9   4  604 [ 0.0/10.9]
```

Figure 3.14: Confusion matrix for read speech: Male

Table 3.5 displays the testing accuracy of PE when trained and tested gender-wise for read speech.

Table 3.5: Testing Accuracy of Read Speech mode – Gender wise

Gender	No. of Speakers	Training Data (in Minutes)	Testing Data (in Minutes)	Testing Accuracy
Female	2	18.48	6.35	52.0%
Male	2	122.41	38.65	61.6%

The PE has also been trained for individual speakers. This has been done for two female speakers and two male speakers. The testing accuracies of this study is depicted in Table 3.6 and Table 3.7.

Table 3.6: Testing Accuracy of read speech mode: Females

Female ID	Training Data (in Minutes)	Testing Data (in Minutes)	Testing Accuracy
F0001	10:35	3.45	50.0%
F0002	7.47	2.4	57.8%

Table 3.7: Testing Accuracy of Read speech mode: Males

Male ID	Training Data (in Minutes)	Testing Data (in Minutes)	Testing Accuracy
M0001	92.33	29.25	61.2%
M0002	29.5	9.98	57.9%

3.3.4.2 Performance Enhancement of Phonetic Engine

To increase the efficiency of the phonetic engine, training as well as testing data sets have been enhanced. The details of the data set that has been used for training and testing the PE is given in Table 3.8.

Table 3.8: Total duration of each mode of speech

Mode	**Duration (in minutes)**
Read speech mode	300.02
Lecture speech mode	40.15
Conversational speech mode	30.02

For read speech, training of PE is divided into 3 steps to measure the correctness and accuracy at each step. In first step, 180 minutes of data has been used, which has further been divided as, 120 minutes data for training and 60 minutes data for testing. It results into 63.2% of accuracy. In second step, 240 minutes of data has been used, which was further divided as 180 minutes data for training, and 60 minutes of data for testing. It resulted into 67.3% accuracy. In third step, 300.02 minutes of data has been used. This data is further divided as 225.02 minutes data for training and 75 minutes data for testing. This approach has resulted into a 71.9% of accuracy.

Table 3.9 depicts the correctness as well as accuracy of PE in incremental manner. As we increase the training data, the correctness and accuracy of PE increases.

Table 3.9: Enhanced performance of read speech mode with increased data sets

Data in hours	**Correctness (in %)**	**Testing accuracy (in %)**
3 hrs	72.2	63.2
4 hrs	75.0	67.3
5 hrs	77.2	71.9

PE is also trained for different number of ASCII symbols. The same incremental approach of data has been applied with different number of ASCII symbols. Experiments show that change in numbers of ASCII symbols affects the performance of PE. The results of these experimentations are depicted in Table 3.10. Results in this Table depict that increased number of ASCII symbols reduce the performance of PE, for same training data sets in comparison to lower number of ASCII symbols.

Table 3.10: Performance of PE with different ASCII symbols

Data in hours	With 29 ASCII symbols		With 49 ASCII symbols	
	Correctness (in %)	Accuracy (in %)	Correctness (in %)	Accuracy (in %)
3 hours	72.2	63.2	49.3	40.6
4 hours	75.1	67.2	51.2	43.6
5 hours	77.3	71.9	53.7	45.1

Phonetic Engine is also trained with read speech data with different number of MFCC features: 12D and 36D MFCC features. The approach of different MFCC features is applied in the incremental manner of data sets and ASCII symbols. The figures on correctness and accuracy is mentioned in Tables 3.11 and 3.12, respectively for these experimentations.

Table 3.11: Correctness of PE with enhanced MFCC features

Data in hours	12D MFCC Features		36D MFCC Features	
	29 ASCII symbols	49 ASCII symbols	29 ASCII symbols	49 ASCII symbols
3 hours	72.2	49.3	72.2	49.3
4 hours	75.1	51.2	75.2	51.2
5 hours	77.3	53.7	78.0	53.9

Table 3.12: Accuracy of PE with enhanced MFCC features

Data in hours	12D MFCC Features		36D MFCC Features	
	29 ASCII symbols	49 ASCII symbols	29 ASCII symbols	49 ASCII symbols
3 hours	63.2	40.6	63.2	40.6
4 hours	67.3	43.6	67.3	43.7
5 hours	71.9	45.1	72.3	45.2

Now, for lecture mode speech, 536 wave files have been considered. Out of these 536 files, 402 files have been used for training purpose and 134 sliced wave files have been used for testing purpose. It resulted into 61.2% of correctness and 50.4% of accuracy, as shown in Figure 3.15.

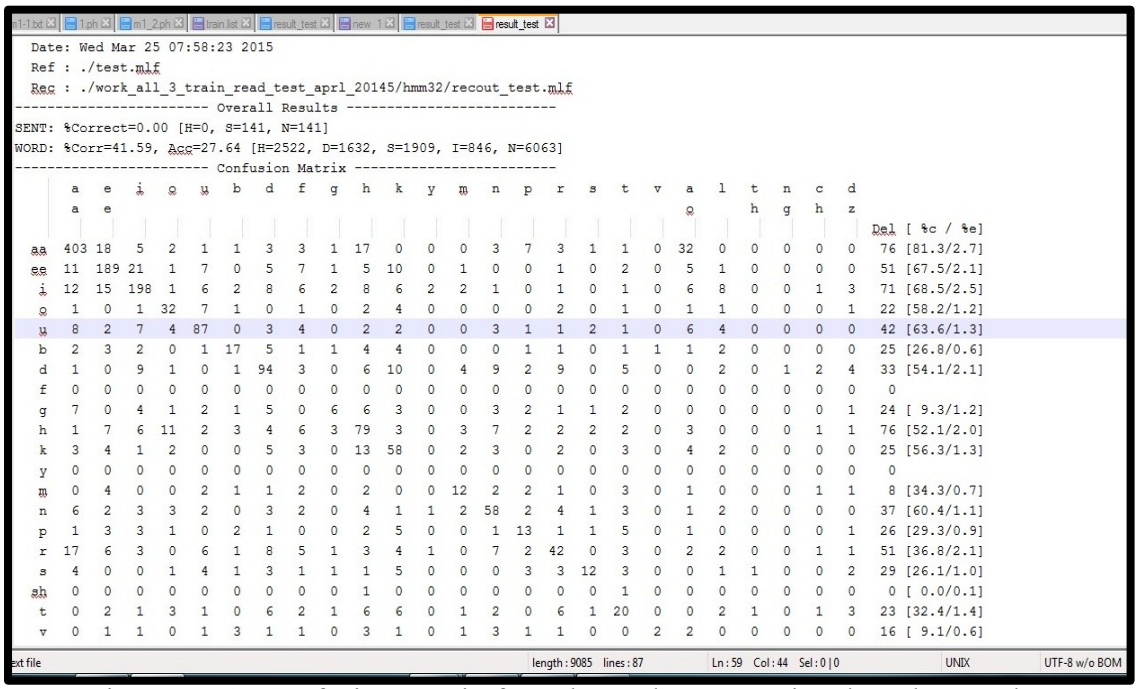
```
Date: Sat Feb 28 11:04:25 2015
Ref : ./test.mlf
Rec : ./work_all_3_train_read_test_aprl_20145/hmm32/recout_test.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=176, N=176]
WORD: %Corr=61.17, Acc=50.41 [H=2932, D=554, S=964, I=421, N=3910]
------------------------ Confusion Matrix -------------------------
```

	aa	ee	i	o	u	b	d	g	h	k	m	n	p	r	s	sh	t	v	a o	l	t h	k h	n g	c h	d z	Del	[%c / %e]
aa	309	11	4	4	1	1	0	0	1	1	1	0	2	4	2	0	4	0	7	0	0	0	0	1	0	26	[89.6/1.5]
ee	21	172	14	0	1	0	1	0	0	0	0	0	1	0	1	0	0	0	2	0	0	0	0	0	0	25	[79.2/1.4]
i	3	19	223	0	3	0	0	0	5	4	1	2	2	1	1	0	2	0	3	1	0	0	0	0	0	29	[85.9/1.5]
o	5	0	0	48	23	0	0	0	2	0	0	0	0	2	1	0	3	2	1	0	0	0	0	0	0	17	[55.2/1.3]
u	6	1	1	8	89	0	0	0	2	2	1	0	1	3	0	0	1	2	3	0	0	0	0	2	0	20	[74.6/1.1]
b	1	0	1	0	1	36	5	0	1	2	2	2	6	0	1	0	3	5	0	0	0	0	0	0	1	18	[47.4/1.1]
d	3	0	0	0	0	0	134	0	0	5	0	4	0	1	0	0	12	1	0	0	0	0	0	0	0	13	[80.7/0.9]
f	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	1	[0.0/0.2]
g	2	1	1	0	0	1	4	8	0	6	0	2	0	3	0	0	3	0	0	1	0	0	0	0	0	12	[29.9/0.9]
h	9	2	6	4	0	1	2	0	28	4	1	3	3	2	1	0	1	0	4	1	0	0	0	0	0	54	[38.4/1.5]
k	0	5	1	0	4	0	1	0	1	67	1	0	8	0	3	0	6	0	0	0	0	0	0	0	2	9	[66.9/1.1]
y	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	[0.0/0.1]
m	0	0	1	0	1	3	3	0	1	1	57	6	1	2	1	0	2	1	0	0	0	0	0	0	0	13	[68.1/0.8]
n	1	0	1	0	2	0	2	0	0	0	8	69	0	3	1	0	0	0	0	0	0	0	0	0	1	10	[83.5/0.6]
p	1	1	0	1	0	0	1	0	1	4	0	0	63	0	0	0	6	0	1	0	0	0	0	0	0	14	[79.7/0.5]
r	7	6	1	2	4	0	4	0	2	2	0	3	1	91	0	0	3	1	3	2	0	0	0	0	1	43	[66.4/1.4]
s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	92	0	1	0	0	0	0	0	0	1	0	12	[97.8/0.1]
sh	0	0	0	0	0	0	0	0	1	0	0	0	0	4	4	0	0	0	0	0	0	0	0	0	1	7	[40.0/0.2]
t	1	1	1	0	1	0	11	0	0	2	0	0	5	0	1	0	80	0	0	0	0	0	0	0	4	21	[72.1/1.0]

```
ext file    length : 9136   lines : 87    Ln : 61   Col : 112   Sel : 0 | 0    UNIX
```

Figure 3.15: Confusion matrix for enhanced lecture mode speech

For conversational mode speech, 410 sliced wave files have been used. Out of these 410 files, 308 files have been used for training purpose whereas 102 files have been used for testing purpose. It resulted into 41.59% of correctness and 27.64% of accuracy, as shown in Figure 3.16.



```
Date: Wed Mar 25 07:58:23 2015
Ref : ./test.mlf
Rec : ./work_all_3_train_read_test_aprl_20145/hmm32/recout_test.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=141, N=141]
WORD: %Corr=41.59, Acc=27.64 [H=2522, D=1632, S=1909, I=846, N=6063]
------------------------ Confusion Matrix -------------------------
```

	aa	ee	i	o	u	b	d	f	g	h	k	y	m	n	p	r	s	t	v	a o	l	t h	n g	c h	d z	Del	[%c / %e]
aa	403	18	5	2	1	1	3	3	1	17	0	0	0	3	7	3	1	1	0	32	0	0	0	0	0	76	[81.3/2.7]
ee	11	189	21	1	7	0	5	7	1	5	10	0	1	0	0	1	0	2	0	5	1	0	0	0	0	51	[67.5/2.1]
i	12	15	198	1	6	2	8	6	2	8	6	2	2	1	0	1	0	1	0	6	8	0	0	1	3	71	[68.5/2.5]
o	1	0	1	32	7	1	0	1	0	2	4	0	0	0	0	2	0	1	0	1	1	0	0	0	1	22	[58.2/1.2]
u	8	2	7	4	87	0	3	4	0	2	2	0	0	3	1	1	2	1	0	6	4	0	0	0	0	42	[63.6/1.3]
b	2	3	2	0	1	17	5	1	1	4	4	0	0	0	1	1	0	1	1	1	2	0	0	0	0	25	[26.8/0.6]
d	1	0	9	1	0	1	94	3	0	6	10	0	4	9	2	9	0	5	0	0	2	0	1	2	4	33	[54.1/2.1]
f	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
g	7	0	4	1	2	1	5	0	6	6	3	0	0	3	2	1	1	2	0	0	0	0	0	0	1	24	[9.3/1.2]
h	1	7	6	11	2	3	4	6	3	79	3	0	3	7	2	2	2	2	0	3	0	0	0	1	1	76	[52.1/2.0]
k	3	4	1	2	0	0	5	3	0	13	58	0	2	3	0	2	0	3	0	4	2	0	0	0	0	25	[56.3/1.3]
y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
m	0	4	0	0	2	1	1	2	0	2	0	0	12	2	2	1	0	3	0	1	0	0	0	1	1	8	[34.3/0.7]
n	6	2	3	3	2	0	3	2	0	4	1	1	2	58	2	4	1	3	0	1	2	0	0	0	0	37	[60.4/1.1]
p	1	3	3	1	0	2	1	0	0	2	5	0	0	1	13	1	1	5	0	1	0	0	0	0	1	26	[29.3/0.9]
r	17	6	3	0	6	1	8	5	1	3	4	1	0	7	2	42	0	3	0	2	2	0	0	1	1	51	[36.8/2.1]
s	4	0	0	1	4	1	3	1	1	1	5	0	0	0	3	3	12	3	0	0	1	1	0	0	2	29	[26.1/1.0]
sh	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	[0.0/0.1]
t	0	2	1	3	1	0	6	2	1	6	6	0	1	2	0	6	1	20	0	0	2	1	0	1	3	23	[32.4/1.4]
v	0	1	1	0	1	3	1	1	0	3	1	0	1	3	1	1	0	0	2	2	0	0	0	0	0	16	[9.1/0.6]

```
ext file    length : 9085   lines : 87    Ln : 59   Col : 44   Sel : 0 | 0    UNIX    UTF-8 w/o BOM
```

Figure 3.16: Confusion matrix for enhanced conversational mode speech

Testing accuracy of each kind of speech mode, corresponding number of speakers, duration of training and testing data for enhanced data set is shown in Table 3.13.

Table 3.13: Testing accuracy of each mode

Mode of Speech	No. of Speakers	Training Data (in Minutes)	Testing Data (in Minutes)	Testing Accuracy
Read	11	225.02	75	71.9%
Lecture	2	30.02	10.08	50.4%
Conversational	10	22.41	7.58	27.6%

3.4 Syllabification Automation

Phonetic engine system for semi-automatic syllabification of audio files is being developed using HTK toolkit. Syllabification is the process of separation of words into syllables. For syllabification, system is trained using 750 wave files. After this, various input files are given as input to toolkit along with HMMs, and a final result is obtained in a file in which duration of each phone of each wav file is obtained. Following figures contain the results of one of such wave file. Figure 3.17 depicts the IPA symbols and their corresponding ASCII characters that have been used in transcription as well as in syllabification. Figure 3.18 depicts the duration of different IPA symbols as a resultant of syllabification process, whereas Figure 19 displays the system generated syllables corresponding to the signal.

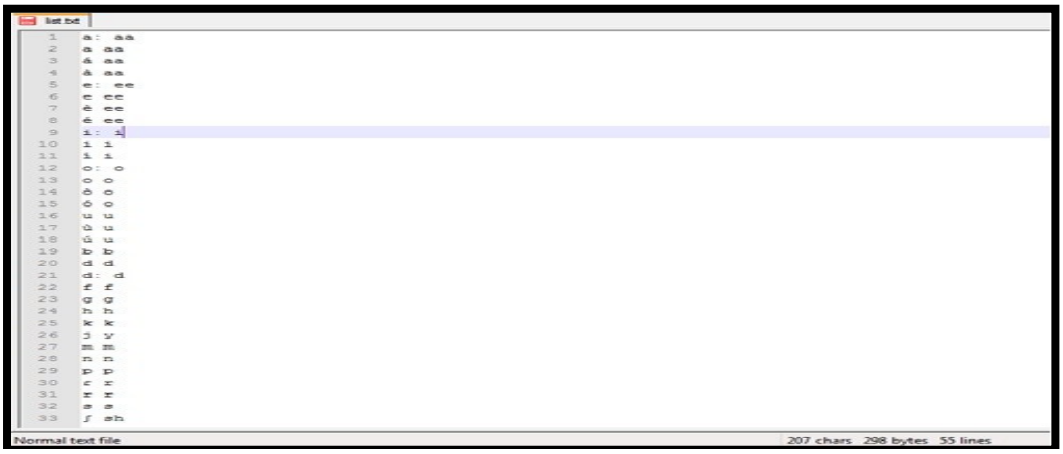


Figure 3.17: IPA symbols and their corresponding ASCII characters

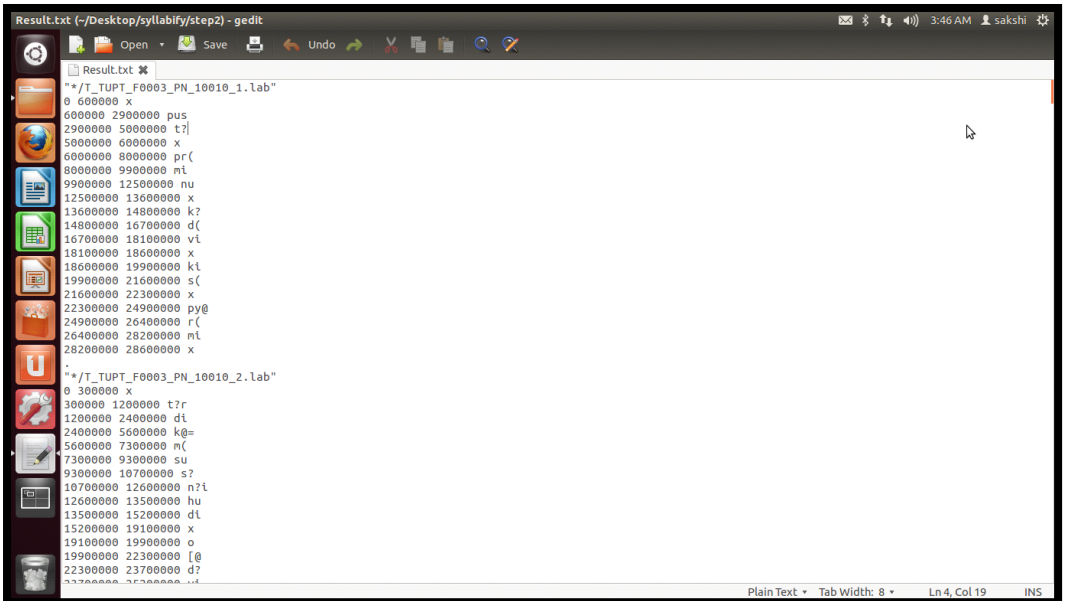


Figure 3.18: Duration of different IPA symbols

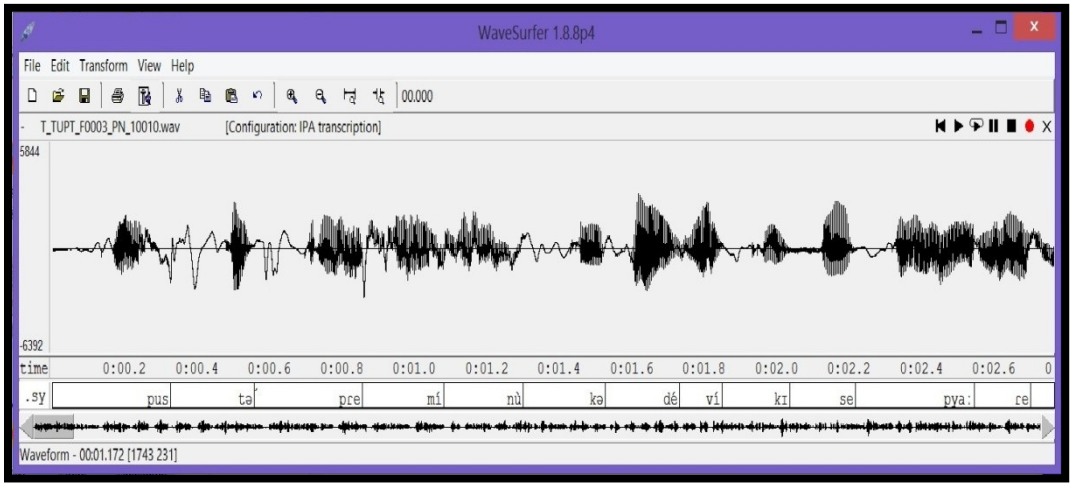


Figure 3.19: System generated syllables corresponding to the signal

3.5 Database Development for searching the speech database

We have prepared the following files for the implementation of search engine for Punjabi Language. These files have been shared with IIT Hyderabad in the form of SPC-1 database. This database contains five directories, namely, convsp, doc, lectsp, readsp, and tools. The directory convsp contains two folders containing conversation speech audio data. The doc folder contains the vocabulary list, query list and word-frequency text files and their pdf files as shown in Table 3.14.

The file vocabulary.txt contains 266 words in total, which are further categorised in mono-syllabic, bi-syllabic, tri-syllabic, and multi-syllabic categories. The occurrences of these syllables in the words is mentioned in Table 3,15.

Table 3.14: Sample vocabulary, word frequency and query.txt

vocabulary.txt	word_frequency.txt	query.txt
ਖੇਤਾਂ ਬੈਠਕ ਵਰਦੀ ਆਦਮੀ ਪਿਸਤੌਲ ਪਿੱਠ ਪੁਲਸੀਆ ਦਿਨੇ-ਰਾਤੀਂ ਇਲਾਕੇ ਡਾਕੂ ਦਿਲ ਮੂੰਹ ਯਾਰ ਬੰਦਾ	0:09.4 0:10.4 ਖੇਤਾਂ 1/2 readsp/m2/1.wav 1:22.2 1:22.8 ਖੇਤਾਂ 2/2 readsp/m2/1.wav 0:18.69 0:18.9 ਬੈਠਕ 1/5 readsp/m2/1.wav 7:13.5 7:14.0 ਬੈਠਕ 2/5 readsp/m2/1.wav 14:06.4 14:06.8 ਬੈਠਕ 3/5 readsp/m2/1.wav 15:36.9 15:37.3 ਬੈਠਕ 4/5 readsp/m2/1.wav 16:20.3 16:21.1 ਬੈਠਕ 5/5 readsp/m2/1.wav 0:19.7 0:20.24 ਵਰਦੀ 1/1 readsp/m2/1.wav 0:20.6 0:21.1 ਆਦਮੀ 1/6 readsp/m2/1.wav 06:26.404 06:26.912 ਆਦਮੀ 2/6 readsp/m1/2.wav 5:29.4 5:29.9 ਆਦਮੀ 3/6 readsp/m1/3.wav 3:33.1 3:33.4 ਆਦਮੀ 4/6 readsp/m1/4.wav	ਬੈਠਕ ਪੁਲਸੀਆ ਇਲਾਕੇ ਡਾਕੂ ਦਿਲ ਮੂੰਹ ਜੈਲਦਾਰੀ ਪੁਲਿਸ ਠਾਣੇ ਸਾਹ ਵਹੁਟੀ ਮੁਲਕਾਂ ਕਿਸਾਨ ਪੰਜਾਬ

Table 3.15: Occurrences of syllables

Words	Occurrences
Mono-syllabic	39
Bi-syllabic	43
Tri-syllabic	110
Multi-syllabic	74

3.6 Summary and future work

In this project, we have collected read mode speech data of 15 hours and 44 minutes duration, lecture mode speech data of 5 hours and 24 minutes duration and conversational mode speech data of 4 hours and 19 minutes duration. A portion of this data has been transcribed for each mode. Experiments have also been carried out for the development of a prosodically guided phonetic engine in this project. The phonetic engines have been developed for different modes of speech. We have also developed a database for implementing the search engine for Punjabi language.

4

IIT Guwahati

PROGRESS SUMMARY REPORT OF IIT Guwahati

A. General

A.1	Name of the Project	:	**Prosodically Guided Phonetic Engine ɑ for Searching Speech Databases in Indian Languages (Assamese)**
	Our Reference Letter No	:	11(6)/2011-HCC(TDIL) dated 23-12-2011
A.2	Executing Agency	:	IIT Guwahati
A.3	Chief Investigator with Designation	:	Prof. S. R. M. Prasanna Professor, Dept. of Electronics and Electrical Engineering
	Co-Chief Investigators with Designation	:	Prof. Samarendra Dandapat Professor, Dept. of Electronics and Electrical Engineering
A.4	Project staff with Qualification	:	Deepak K. T. (PhD Student, EEE Dept.), Biswajit Dev Sarma (PhD Student, EEE Dept.), Mousmita Sarma (M.Tech.(ECT), GU), Meghamallika Sarma (B.E.(AI), GU), Abhishek Dey (B.E.(ECE), GU)
A.5	Total Cost of the Project as approved by DIT	:	
	i) Original	:	40.25 Lakhs
	ii) Revised, if any	:	–
A.6	Date of starting (Indicate date of first sanction)	:	23/12/2011
A.7	Date of Completion	:	
	i) Original	:	
	ii) Revised, if any	:	
A.8	Date on which last progress report was Submitted	:	28-02-2014

A.9 Work in Progress (Details are given in technical report in Appendix 4.1)

Appendix 4.1

Detailed Technical Report of IIT Guwahati

A Comprehensive Report on Development of Phonetic Engine for Indian Languages

Detailed Technical Report of IIT Guwahati

4.1 Database collection & transcription

4.1.1 Data Collection

Data has been collected from around twenty five native Assamese speakers for a duration of about 20 hours. Recording has been done in three different modes. The environment chosen during the lecture mode data collection process while recording is the normal academic class environment. For reading mode it is open or closed room environment. Conversation mode is collected in field environment. Two channels have been used for recording the data for lecture and reading mode. One channel is the narrowband telephone channel and the other channel is the microphone channel. Microphone channel is maintained at a sampling frequency of 48 KHz, whereas sampling frequency of the narrowband telephone channel is 8 KHz. Both channels are maintained at 16 bit per sample. Asterisk voiceserver is used to record the telephone data. Only mobile channel is used for conversational mode dat collection.

4.1.2 Transcription

After the commencement of the recording process, the data obtained needs to be chunked and transcribed. Chunking of the data into smaller parts has been done comparable to the length of a general sentence. Once chunking gets over, the data is ready for being transcribed. While transcribing, the signal is carefully listened and looked into so as to minimize transcription error as much as possible. Transcription has been done using the International Phonetic Alphabet (IPA) chart. The International Phonetic Alphabet (IPA) is an alphabetic system of phonetic notation based primarily on the Latin alphabet. The IPA is designed to represent those qualities of speech that are distinctive in oral language. It provides one symbol for each distinctive sound (speech segment) composed of

one or more elements of two basic types, letters and diacritics. Letters represent basic sound units and diacritics are small markings which are placed around the IPA letter in order to show a certain alteration or more specific description in the letters pronunciation. Since the IPA symbols captures all the distinctive acoustic phonetic characteristics of speech, they are used as phonetic units in the proposed PE. The most recent version of the IPA chart consist of 107 letters, 52 diacritics and four prosodic marks. Transcribers are free to use any one of the 107 letters from the chart, however, diacritics set for transcription is reduced to 10 containing only the common ones like aspirated stops, nasalised sounds, long vowel, extra short vowel, half long vowel, breathy voice etc. However, no prosodic marking is done.

Segmented speech files are carefully listened to and observed to transcribe using the phonetic units obtained from the IPA chart. WaveSurfer is used to listen and visually examine the speech waveform. To remove ambiguities among the phonetic units, they are observed using different signal processing tools like spectrogram, pitch contour, energy contour etc. Whole 20 hours of speech data is transcribed using IPA. Around 2 hours of data is marked with prosodic markings (syllable marking, pitch contour marking and break index marking). A total of about 70 different phonetic units are found for the Assamese data. However, most of the phonetic units with diacritics are found to have very low number of occurences and so they are merged to the corresponding phonetic unit without the diacritic mark. Also, most of the phonetic units which are not commonly found in Assamese but produced by the articulators in continuous speech have very less number of occurences. Number of occurrences of 35 major phonetic units from 3.5 hours of data are tabulated bellow:

Vowels	ɒ	ɔ	i	a	ɛ	o	e	u
No. of Occurrence	90	8646	4329	5814	3132	1502	637	3343

Nasals	n	m	ŋ
No. of Occurrence	1843	296	233

Unaspirated stops	p	b	t	d	k	g
No. of Occurrence	1522	2166	4225	1014	3815	469

Aspirated stops	ph	bh	th	dh	kh	gh
No. of Occurrence	197	402	539	265	49	541

Fricatives	z	s	x	h	β	χ	ʃ
No. of Occurrence	243	2314	1359	1478	57	44	30

Approximants	r	j	l
No. of Occurrence	5055	827	2356

Affricates	ts	dz
No. of Occurrence	54	502

4.2 Semi-automatic prosodic markings

4.2.1 Pitch Contour marking

In a particular voiced segment of speech, pitch may vary from low to high, high to low or it may not vary at all. This work describes a method for automatically segmenting speech into certain regions having a continuous pitch contour and marking the nature of pitch change within those regions. Zero frequency filtering is used to segment the speech into voiced and unvoiced segments. This segment is further divided into small segments depending on a discontinuity present in the pitch contour. A height value of the pitch contour in the final segment is measured and accordingly marking is done. Now automatic segmentation and markings are manually corrected by deleting, inserting or shifting the segmentation boundaries and substituting the wrong markings.

Hundred speech sentences containing around 700 pitch contour segments collected from 30 different speakers are used for evaluation of automatic segmentation and marking of pitch contours. Performance is evaluated in terms of percentage of manual error correction.

- DEL: Percentage of deleted segment boundaries in the manual correction process out of total number of actual boundaries.

- INS: Percentage of inserted segment boundaries in the manual correction process out of total

number of actual boundaries.

- SFT: Percentage of shifed segment boundaries in the manual correction process out of total number of actual boundaries.

- SUB: Percentage of substituted pitch contour markings in the manual correction process out of total number of actual pitch markings.

Table 4.1: Performance evaluation

DEL (%)	INS (%)	SFT (%)	SUB (%)
9.38	7.18	6.89	18.91

4.2.2 Semi-automatic syllable boundary marking

Semi-automatic syllable labelling means syllable labelling of the speech signal when transcription or the text corresponding to the speech file is provided. HMM models for 15 broad classes of phone is built. Time label of the transcription is obtained by the forced alignment procedure using the HMM models. A parser is used to convert the word transcription to syllable transcription using certain syllabification rules. This syllable transcription and the time label of the phones are used to get the time label of the syllables. Now the syllable labelling output is refined using the knowledge of vowel onset point and vowel offset point derived from the speech signal using different signal processing techniques. This refinement gives improvement in terms of both syllable detection as well as average deviation in the syllable onset and offset.

50 sentences from the Assamese database which are not part of the training data are taken and manually labelled the syllables using wave-surfer. Pitch contour, spectrogram etc are used for accurate labelling. Now the same sentences are tested with the proposed system. The performance is evaluated using the following parameters.

Detection Rate (DR): Percentage of syllable onset and offset detected within 40 ms deviation of actual syllable onset and offset.

Average Deviation (AD): Average of the deviations of all the detected syllable onsets and offsets.

Spurious detection (SD): Syllable Onset or Offset that is detected beyond 40 ms of an actual Onset or Offset.

Table 4.2: Syllable labelling performance

Method	DR (%)	AD (msec)	SD (%)
HMM	92.81	14	7.19
HMM+VOP/VEP	93.35	12	6.65

Table 6.4 shows the performance of the system before and after refinement. It is seen that using only HMM based method gives 92.81% detection rate which is improved to 93.35% after refining with the knowledge of VOP and VEP. In case of average deviation, HMM gives around 14 msec of average deviation, whereas, the refined output gives 12 msec of average deviation. Thus, after refinement, improvement is observed in terms of detection rate and average deviation.

4.3 Development of phonetic engine

Phonetic engine (PE) is the signal to symbol transformation module which uses the acoustic phonetic information present in the speech signal to convert the speech signal into symbolic form. The engine produces a sequence of symbols without using any language constraints in the form of lexical, syntactic and higher level knowledge source. The choice of symbol should be such that it can capture all the phonetic variations in speech. Existing PE implemented for Indian languages produces syllable like units as the output where constraint at the syllable level are used, as syllable-like units are most basic in the production of speech.

PE is the front end module for both speech recognition system and information retrieval system. In automatic speech recognition of continuous speech, the speech signal is first converted to the subword units of speech which in turn is converted to text. The first part of converting speech to subword units is done by a PE. Existing PE implemented for Indian languages uses syllable like units as the subword units. Here we will use sequence of International Phonetic Alphabet (IPA) as the subword units as IPA provides one symbol for each distinctive sound (speech segment). These symbols are composed of one or more elements of two basic types, letters and diacritics. Letters represent basic sound units and Diacritics are small markings which are placed around the IPA letter in order to show a certain alteration or more specific description in the letter's pronunciation. Since IPA symbol captures all distinctive acoustic phonetic characteristics of speech, they can be called as acoustic phonetic sequence (APS). In an information retrieval system, the spoken query is converted to Acous-

tic APS by PE. The APS derived from PE is fed as an input to the IR module, where the job of IR module is to use APS to search and retrieve the relevant spoken documents based on the spoken query.

4.3.1 Signal Processing based approach

Here we try to detect the phonetic units in a hierarchical way from a very high level segmentation to a low level segmentation. First, speech signal is classified into vowel like regions and non-vowel like regions. Vowel regions includes vowels and semivowels and non vowel like regions includes nasals, stops and fricatives. Later, vowel like regions and non vowel like region will be recognised using separate methods. Here we use a excitation source based method for detection of vowel like regions. The output of the signal processing based method is refined using a HMM based statistical method. Signal processing based method for vowel like region detection uses hilbert envelope of LP residual of speech signal and strength of excitation of speech derived from zero frequency filtered signal. VLR detection using SP method mainly relies on the change of signal strength. Nasals are quasi-periodic in nature with impulse like excitation source characteristics in it. So, there can be a change in strength at the onset of a nasal. Similarly voice bars, having high energy can be detected as VLR as they have the similar excitation source characteristics. However, vocal tract information present in the statistical method is capable of detecting such regions. A multi-class statistical phone classifier that classifies speech into broad vowel, consonant and silence categories is trained. The outputs of the classifier are suitably combined to get evidence for vowel-like regions, different broad categories of consonants and silence regions. The output from the existing signal processing method is compared with different evidences from the statistical method. The spurious ones are eliminated by using the evidences from the statistical method. The experimental studies conducted on TIMIT and in-house databases demonstrate significant reduction in the spurious VLRs with a little loss in the VLRs detection performance. A net gain of 4.21% and 7.71% in frame error rate is achieved for TIMIT and in-house databases, respectively.

Another method for improving accuracy of Vowel Onset Point (VOP) and Vowel End Point (VEP) detection in continuous speech is devloped. Speech signal is represented using Bessel functions with their damped sinusoid-like basis functions. Bessel expansion is used to emphasize the vowel regions by appropriate consideration of the range of Bessel coefficients. Bandpass filtered narrow-band signal

Table 4.3: VLRs Detection Performance after reducing Spurious detection rate

Data	Method	IR(%)	SR1(%)	SR2(%)	NG(%)
TIMIT	SP	80.38	12.40	0.81	-
	$SP - ST_N$	77.65	8.37	0.79	1.33
	$SP - ST_S$	78.34	7.38	0.49	3.35
	$SP - ST_F$	80.43	12.44	0.81	0.02
	$SP - ST_S - ST_N - ST_F$	74.75	2.92	0.46	4.21
Assamese	SP	88.32	22.48	2.46	-
	$SP - ST_N$	85.15	18.01	2.42	1.20
	$SP - ST_S$	82.03	11.07	1.77	5.67
	$SP - ST_F$	87.34	21.01	2.23	0.58
	$SP - ST_S - ST_N - ST_F$	77.80	5.01	1.48	7.79

is modeled as a monocomponent amplitude modulated-frequency modulated (AM-FM) signal. The amplitude envelope (AE) function of this vowel emphasized AM-FM signal gives strong evidence for the VOP and VEP. This evidence after adding with some of the existing evidences having source and system information, increases the detection rate as well as the accuracy of detection. Evidences obtained using Excitation source (ES) and source, spectral peaks and modulation spectrum (SSM) based methods are enhanced in this work by adding the evidence obtained from the AE function of the vowel enhanced signal. Normally, the evidence from AE function will have a strong peak at the VOP and VEP compared to other speech region within the same vowel. Adding this evidence will enhance the ES and SSM evidence at the VOP and VEP. Even if the peaks at VOP or VEP are not strong enough, but almost comparable, in the combined evidence, the peaks will move towards the VOP or VEP. After adding the evidences same procedure is followed for the respective methods for obtaining the VOP or VEP.

Table 4.12 shows the performance of VOP/VEP detection in terms of DR and SR. Individual performances of ES and SSM are compared with corresponding combined performances (AE+ES and AE+SSM). Improvement is achieved in terms of both DR and SR in terms of increasing DR and reducing SR.

4.3.2 HMM based approach

Diacritics are found to have very low number of occurences and so they are merged to the corresponding phonetic unit without the diacritic mark. Most of the phonetic units which are not commonly

Table 4.4: VOP/VEP Detection Performance

Method	VOP		VEP	
	DR (%)	SR (%)	DR (%)	SR (%)
ES	94.06	8.17	92.14	10.09
ES+AE	95.33	6.63	92.95	8.82
SSM	93.56	9.03	87.21	14.76
SSM+AE	95.12	7.64	89.31	12.87

found in Assamese but produced by the articulators in continuous speech have very less number of occurences. Such phonetic units are merged to the closest IPA symbol. After merging the phonetic units a total of 34 symbols are found including the silence.

4.4 Development of Phonetic Engine for 12 Indian Languages

The PEs for 12 Indian Languages are developed in this work using Hidden Markov Model (HMM) and Artificial Neural Network (ANN) approach. The speech corpus for each language consists of phonetically transcribed speech data in different speaking modes i.e. read, lecture and conversation. The phone level transcription is done by using IPA symbols and diacritic. Since PEs are developed by statistical methods, requires more examples to get better trained models. Hence in each language merging of phonetic units to enhance the number of examples for each phonetic class is carried out in this work. Diacritics are found to have very low number of occurrences and so they are merged to the corresponding phonetic unit without the diacritic mark. Most of the phonetic units which are not commonly found in the languages, but produced by the articulators in continuous speech have very less number of occurrences. Such phonetic units are merged to the closest IPA symbol. The number final merged phonemes used to develop PEs for different Indian languages is given below.

Mel Frequency Cepstral Coefficients (MFCCs) are computed for Hamming windowed speech frames of 25ms with 10ms overlap. In this work, 22 Mel scaled filter banks are used compute 12^{th} order MFCCs. Hence, speech is parameterized with 12 Mel Frequency Cepstral Coefficients (MFCCs) and 0^{th} order cepstral coefficient as well as their first and second order derivatives with zero mean static coefficient yielding a total of 39 components. The same set of feature vectors are used to develop the HMM and ANN based PEs.

4.4.1 HMM Based Approach

Context-independent mono-phone hidden Markov model (HMM) based phone recognizer is used to build the prototype of the phonetic engine. Here phones are the acoustic phonetic units (APU) which are prepared using IPA. The system will not use language information in any form. A 5 states left to right HMM model (including 2 non emitting states) with a 32 mixture continuous-density diagonal-covariance Gaussian mixture model (GMM) per state will be used to model each of the phonetic units. The training process is initialized by defining prototype model, where the model topology such as type of features used, dimension of feature vector and number of states and transition probabilities are defined. In this work, a prototype model with all means initialized to zeros and all variances initialized to unity is defined. The global means and variances are computed by scaning all the training files. A new prototype model with all its means equal to global means and all its variances equal to global

variances, for set of all Gaussians present in the given HMM, is created. This new prototype model is used for creating flat-start HMMs. The flat start HMMs are then re-estimated using the training data with the embedded re- estimation to perform Baum-Welch training. Here 6 iterations are used to get final phone models.

The viterbi decoding is used for decoding of test speech signal into sequence of phones. Viterbi decoding is a procedure for finding the hidden sequence of states within a phone. These states are most likely to have produced the observed sequence of feature vectors.

The phone recognition accuracy is computed by using,

$$RecognitionAccuracy(in\%) = \frac{N - S - D - I}{N} \tag{4.1}$$

Where, N is total No. of phones, S is No. of Substitutions, D is No. of Deletions and I is No. of Insertions.

4.4.2 ANN Based Approach

The training of ANN requires phonetically transcribed speech data with time stamps. Since ANN is discriminative classifier, it requires phone boundary information for training. The mono phone HMMs are used to conduct force alignment on the training data to get the time stamps. The phones with time stamps are used to train the ANN. The input layer of ANN is equal of the feature vector, i.e. 39 if 0 context is used. Since the phone boundary information is taken from a statistical model (HMM) accurate phone boundaries cannot be expected. Context of order 8 is used to obtain the input feature vector for ANN i.e. feature vectors of current frame and 4 frames on either side current frame. The size of output layer equals to the number of phones used. The hidden layer of size 1000 is used. Back propagation algorithm is used train the Feed Forward ANN (FANN). The structure of FANN used to develop the phonetic engine in the current work is [351 1000 Number of Phones]. Fig. 4.1 shows the structure of Phonetic Engine developed using ANN. The phone decoding of testing sequence is done by decoding posteriori probabilities of output layer of ANN.

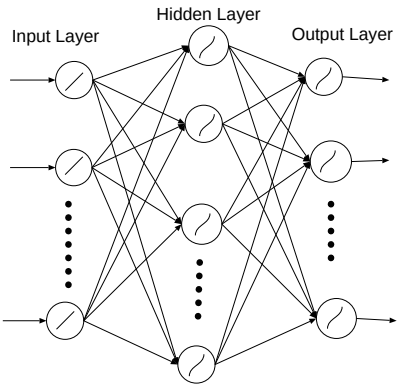


Figure 4.1: The Architecture of ANN based phonetic Engine. In present work ANN of structure ([351 1000 No. of Phones]), input layer of size 351, hidden layer of 1000 neurons and output layer of size equals to number of phones are used

4.4.3 Phonetic Engine Results

The accuracies of phonetic engines developed for 12 languages with three different speaking modes are shown in table 4.7. The ANN based phonetic engine shows better performance compared to HMM based phonetic engine, because of discriminative training of ANN system. The relative increment in performance of ANN w.r.t. HMM is also indicated in table 4.7. The experiments on different modes of training and testing data are carried out using HMM, to study the effect of speaking mode on the performance of phonetic engines. The results of these experiments are shown in Appendix-I.

4.5 Computation of Confidence Level for the Decoded Phone Sequence

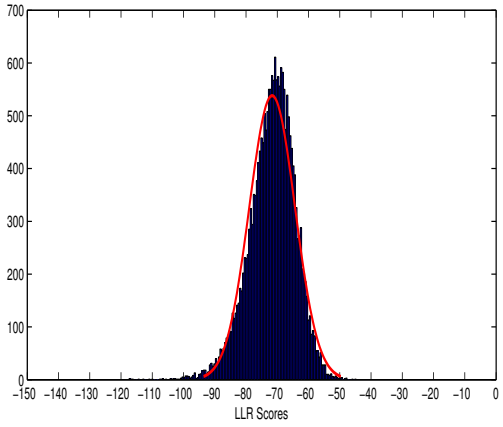


Figure 4.2: Histogram of LLR scores of phone [a]

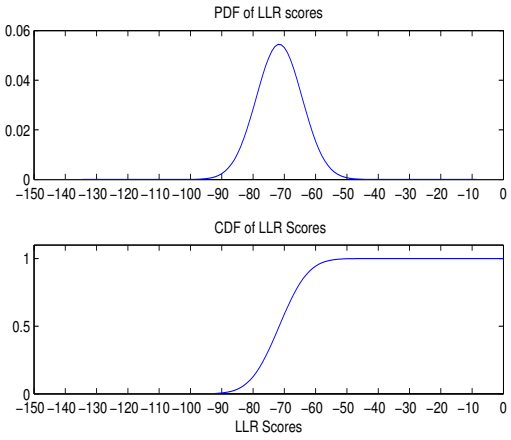


Figure 4.3: The Gaussian PDF (a) and CDF (b) of scores of phone [a].

The phonetic engine decodes the given utterance into a sequence of phones. The confidence level of decoded phone is obtained using output score of PEs. The forced alignment is conducted on the train data and corresponding LLR scores for the forced aligned phones is obtained. The phone wise LLR scores are grouped together and distribution of scores is computed. The histogram of LLR scores of phone [a] of Kannada language is shown in Fig. 4.2. The distribution of decoded phone scores follows the Gaussian Probability Distribution Function (PDF) . So, LLR scores are modeled by Gaussian PDF Fig. 4.3. The corresponding Cumulative Distribution Function (CDF) is computed from the PDF obtained for a particular phone. The Fig. 4.4 shows the computation method of CDF contours for the confidence level measurement of decoded phones of phonetic engine. The table 4.8 shows the CDF contours computed for the LLR scores. The CDF distributions are computed for closed interval [-150,0] with bin size of 0.01.

During testing, the phonetic engine decodes the phone sequence with LLR scores. The CDF value corresponding to the LLR score gives the confidence level of the decoded phone. The Fig. 4.5 explains the computation of confidence level for the decoded phone sequence from the CDF contours computed. An example of computing confidence scores for the decoded phones of utterance “akashavani” is shown in table 4.9. The correct decoded sequence of the utterance will be [a], [k], [a], [sh], [a], [v], [a], [n], [i], but decoded as [a], [t], [k], [a], [s], [e], [v], [a], [n], [i], [T]. The decoded phones [a] (0.60),[a] (0.67), [v]

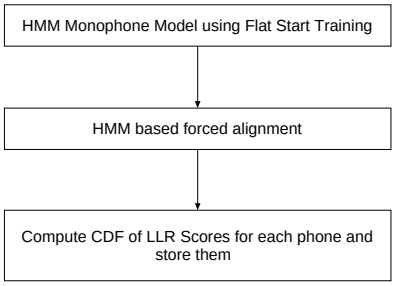


Figure 4.4: Computation of CDF for the LLR scores of training data.

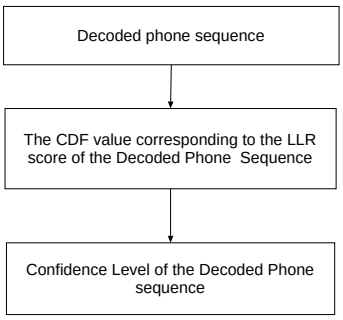


Figure 4.5: Computation of confidence level for the test utterance from the CDF obtained from LLR scores.

(0.99), [a] (0.99) with high confidence value shows the presence of decoded phones in the utterance.

4.6 Language Identification using Phonetic Engine

The influence of language on the phone recognition is illustrated by decoding a given utterance in all the 12 language phonetic engines. The given utterance is decoded in phonetic engines of all the 12 languages and corresponding LLR scores are computed. The phonetic engine which gives the maximum score for the utterance is identified as the language of the given utterance. The pictorial representation of language identification is shown in Fig. 4.6.

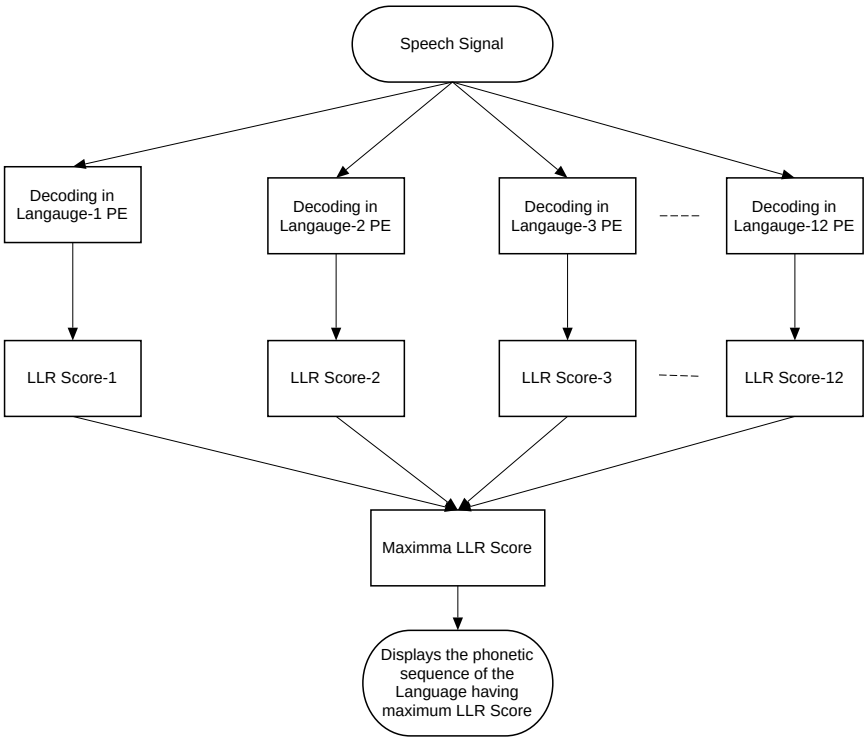


Figure 4.6: Language Identification using Phonetic Engine. The phone decoding of the given utterance is based on the maximum LLR score obtained from all 12 language phonetic engines.

The table 4.10, 4.11 and 4.12 gives the confusion of phone decoding among 12 languages in read, lecture and conversation modes respectively. The diagonal values of the matrix are high compared to others, which shows that the given utterance is decoded in the corresponding language phonetic engine compared to others. The experimental results shows that the maximum number of utterances are decoded in the corresponding languages with maximum score compared to other languages. This shows the influence of language on the phonetic engine performance.

4.7 Phonetic Engine Graphical User Interface (GUI)

The utterance is loaded to decode the phone sequence. The decoding of the phonetic sequence is based upon the user input which may be Language Independent or Language Dependent. The GUI is developed such that the user is free chose language, mode and HMM or ANN based decoder.

If the user selects the Language Independent option, then he has to select the mode and type of engine i.e. HMM or ANN. The utterance is decoded based upon the Language Identification method i.e. decoded based by computing maximum likelihood score among 12 language decoders. If the user selects the Language Dependent option, then he has to select the language, decode and mode for which he wants the sequence to be decoded. The pictorial representation of the GUI is shown in Fig. 4.7.

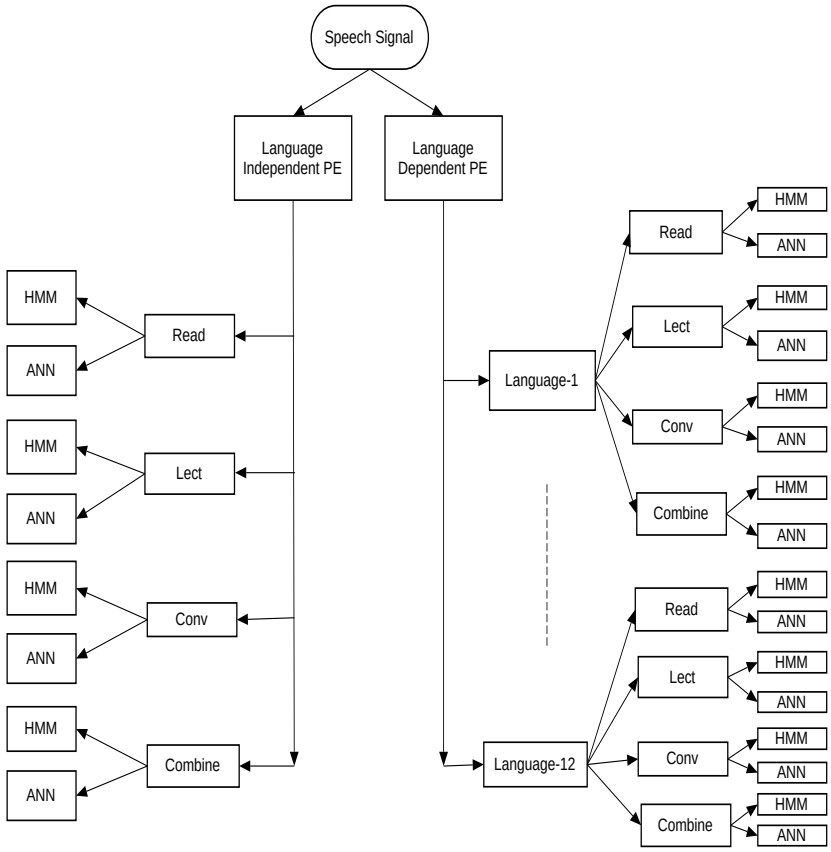


Figure 4.7: Schematic Representation of GUI. The figure explains the flow of decoding of given utterence

The display part of the GUI consists of four layers as shown in figure 4.8. The first layer displays of waveform of the speech signal which has to be decoded into a set of phonetic units. The second layer displays the decoded phone sequence by HMM or ANN. The third layer contains the phones

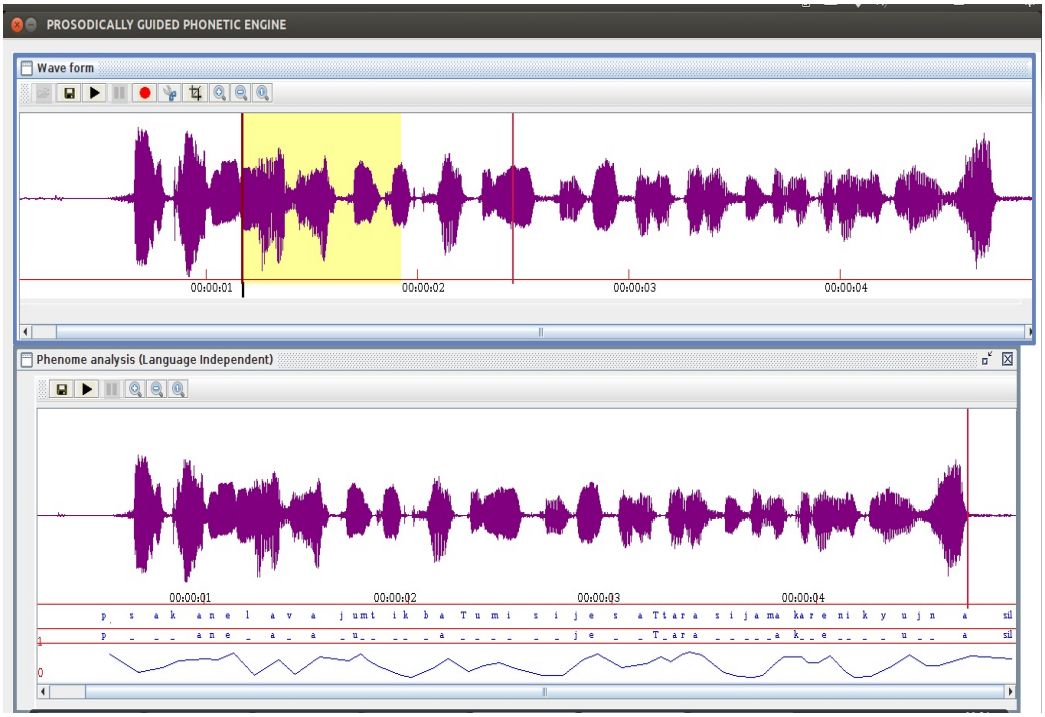


Figure 4.8: Snapshot of GUI for Phonetic Engine. The figure shows the waveform and representation of decoded phone sequence in four different layers in GUI

decoded with high confidence score (¿0.6) and phones with confidence score (¡0.6) are indicated by symbol []. The fourth layer displays the contour of confidence level of decoded phones. The confidence level contour help the user to read the decoded phone sequence in better way i.e. phones with high confidence level given more importance compared to that of less confidence level.

4.8 Conclusion

The development of phonetic engine for 12 Indian languages is exaplained in this work. The merging of phones to create sufficient amount of examples for the training of modeles is done for all languages. Hence the phonetic engines developed for each language decodes the utterence into a sequence of phonetic units with varying no. of phones i.e. 25 to 38 (depending on the language). The

GUI provides the flexibility to user to decode the utterence in any language, mode or PE (HMM or ANN) of his choice. The confidence scores of decoded phones indicates the presence or absence of decoded phone in the given utterence. The future work may includes the incorporation of acoustic-phonetic features to develop the phonetic engine.

4.9 Papers Published Related to the Project Work

[1] Biswajit Dev Sarma and S. R. Mahadeva Prasanna, "Analysis of Vocal Tract Constrictions using Zero Frequency Filtering", IEEE Signal Processing Letters, vol. 21, no. 12, Dec 2014.

[2] Biswajit Dev Sarma, Mousmita Sarma, S R M Prasanna,"Semi-automatic Syllable Labelling for Assamese language using HMM and Vowel onset-offset points", in Proc. WACC 2014, Guwahati.

[3] Biswajit Dev Sarma, Meghamallika Sarma, S R M Prasanna, "Semi-automatic Segmentation and Marking of Pitch Contours for Prosodic Analysis",in Proc. WACC 2014, Guwahati.

[4] Biswajit Dev Sarma, Supreeth Prajwal S and S. R. Mahadeva Prasanna, "Improved Vowel Onset and Offset Points Detection Using Bessel Features" in Proc. International conference on signal processing and communication, 2014, IISc Bangalore.

[5] Biswajit Dev Sarma and S. R. Mahadeva Prasanna, "Analysis of spurious Vowel like regions (VLRs) detected by excitation source information"in Proc. Indicon, Dec. 2013, IIT Bombay.

[6] Biswajit Dev Sarma, Mousmita Sarma, Meghamallika Sarma and S. R. Mahadeva Prasanna, "Developement of Assamese phonetic Engine: Some issues"in Proc. Indicon, Dec. 2013, IIT Bombay.

[7] K. T. Deepak and S. R. Mahadeva Prasanna, "Remote Spoken Document Retrieval using Foreground Speech Segmentation based Isolated Word Recognizer," in Proc Indicon, Dec. 2013, IIT Bombay.

Table 4.5: List of Phonetic units found at the end of transcription of the Assamese data and the Reduced set after merging similar units

Phonetic units	Reduced Phonetic units	Name in ASCII
ɔ,ɒ	ɔ	ao
i, ɪ	i	i
a, ə, ɑ	a	aa
ɛ	ɛ	e
o	o	o
e	e	ee
u, ʊ	u	u
n	n	n
m	m	m
ŋ	ŋ	ng
p	p	p
b	b	b
t	t	t
d	d	d
k	k	k
g	g	g
p^h	p^h	ph
b^h	b^h	bh
t^h	t^h	th
d^h	d^h	dh
k^h	k^h	kh
g^h	g^h	gh
z	z	j
s	s	s
x, ɕ, ɯ	x	x
h	h	h
w, β	w	w
ʃ	ʃ	sh
ɹ, ɾ	ɹ	r
j	j	y
l	l	l
ts	ts	ts
dz	dz	dz

Table 4.6: The number of phones used to build PE for 12 Indian Languages

Sl. No.	Language	No. of phones
1	Assamese	34
2	Bengali	34
3	Gujarati	37
4	Hindi	38
5	Kannada	26
6	Malayalam	25
7	Manipuri	30
8	Marathi	37
9	Odia	34
10	Punjabi	30
11	Telugu	25
12	Urdu	25

Table 4.7: The HMM and ANN based Phonetic Engine Results. Table Shows the training and testing duration, No. of speakers in training and testing and performance HMM and ANN based phonetic Engines for all 12 Languages with 3 different speaking modes

Language	Train Mode	Test Mode	Train Duration (Hrs:Mins:Secs)	Test Duration (Hrs:Mins:Secs)	No. of Training Speakers	No. of Testing Speakers	Accuracy in %	
							HMM	ANN
Manipuri	Read	Read	03:28:26	01:31:59	9 (3M+6F)	6 (2M+4F)	62.11	68.70
Manipuri	Lect	Lect	01:39:44	00:50:53	4 (2M+2F)	2 (1M+1F)	59.78	57.44
Manipuri	Conv	Conv	02:03:56	00:27:21	2 (2M)	2 (2M)	53.92	56.79
Assamese	Read	Read	06:16:20	02:31:20	14 (9M+5F)	6 (6F)	49.60	62.95
Assamese	Lect	Lect	03:34:00	00:52:23	4 (1M+3F)	2 (2F)	52.26	53.10
Assamese	Conv	Conv	02:01:39	00:31:17	8 (3M+5F)	4 (4M)	29.82	35.07
Odia	Read	Read	03:46:37	01:19:24	22 (10M+12F)	9 (4M+5F)	59.69	72.09
Odia	Lect	Lect	01:51:21	00:37:29	6 (3M+3F)	2 (1M+1F)	43.43	56.10
Odia	Conv	Conv	02:13:07	00:20:41	6 (3M+3F)	2 (1M+1F)	49.58	61.25
Bengali	Read	Read	03:49:55	01:10:55	38 (14M+24F)	12 (6M+6F)	43.23	55.47
Bengali	Lect	Lect	01:59:29	00:31:14	7 (5M+2F)	4 (2M+2F)	35.80	36.47
Bengali	Conv	Conv	01:59:11	00:38:31	24 (19M+5F)	6 (3M+3F)	23.99	29.61
Telugu	Read	Read	03:27:19	01:11:12	11 (5M+6F)	7 (3M+4F)	56.08	59.82
Telugu	Lect	Lect	01:29:45	00:30:59	7 (6M+1F)	2 (1M+1F)	40.20	38.49
Telugu	Conv	Conv	01:46:51	00:39:01	34(30M+4F)	13 (10M+3F)	45.83	48.30
Urdu	Read	Read	03:34:50	01:12:13	32 (29M+3F)	25 (22M+3F)	45.67	44.86
Urdu	Lect	Lect	01:35:39	00:36:08	3 (2M+1F)	1 (1M)	54.83	45.47
Urdu	Conv	Conv	01:42:49	00:31:06	16 (8M+8F)	7 (4M+3F)	48.84	47.42
Marathi	Read	Read	05:25:41	02:11:54	61 (42M+19F)	23 (18M+5F)	52.82	56.93
Marathi	Lect	Lect	02:06:51	00:42:16	7 (3M+4F)	2 (1M+1F)	35.18	39.66
Marathi	Conv	Conv	01:41:57	00:35:31	26 (19M+7F)	10 (4M+6F)	35.61	39.81
Gujarati	Read	Read	04:09:33	01:24:55	95 (54M+41F)	28 (17M+11F)	63.55	65.44
Gujarati	Lect	Lect	02:16:00	00:40:26	5 (4M+1F)	4 (3M+1F)	51.52	49.47
Gujarati	Conv	Conv	03:12:07	01:05:35	72 (37M+35F)	46 (26M+20F)	55.35	57.37
Pujabi	Read	Read	02:18:12	00:47:44	3 (1M+2F)	2 (2M)	61.48	74.57
Pujabi	Lect	Lect	00:15:18	00:04:45	1 (1M)	1 (1M)	47.17	50.96
Pujabi	Conv	Conv	00:15:06	00:05:07	10 (9M+1F)	5 (4M+1F)	22.39	28.81
Malayalam	Read	Read	03:58:50	01:13:57	22 (15M+7F)	8 (3M+5F)	33.23	39.80
Malayalam	Lect	Lect	01:56:11	00:35:41	13 (10M+3F)	6 (3M+3F)	31.82	34.82
Malayalam	Conv	Conv	01:52:59	00:37:19	41 (31M+10F)	8 (3M+5F)	31.18	32.37
Kannada	Read	Read	01:56:06	00.29.40	11 (5M+6F)	2 (1M+1F)	55.10	60.11
Kannada	Lect	Lect	01:56:11	00.39.14	4 (3M+1F)	2 (1M+1F)	48.74	53.33
Kannada	Conv	Conv	01:52:59	00.27.43	12 (11M+1F)	6 (6M)	45.67	50.78
Hindi	Read	Read	02:09:17	00:44:54	29 (7M+22F)	14 (8M+6F)	49.24	48.60
Hindi	Lect	Lect	00:43:28	00:17:15	9 (8M+1F)	8 (8F)	37.96	34.71
Hindi	Conv	Conv	02:00:09	00:32:44	18 conv	9 (5M+4F)	42.15	37.41

Table 4.8: Example of Confidence Level Contour

LLR Scores									
Phones	-150	-149.99	-149.98	...	-60.99	-61	..	-0.01	0
[a]	0	0	0	...	0.66	0.66	...	1	1
[e]	0	0	0	...	0.77	0.78	...	1	1
[i]	0	0	0	...	0.54	0.55	...	1	1
[o]	0	0	0	...	0.67	0.78	...	1	1
[u]	0	0	0	...	0.78	0.79	...	1	1

Table 4.9: Phone level decoding of Kannada utterance “akashavani“ with confidence levels

Decoded Phone	LLR score	Confidence level
[a]	-79.03	0.600797
[t]	-78.74	0.641425
[k]	-87.83	0.090313
[a]	-78.12	0.675898
[s]	-86.45	0.037411
[e]	-84.76	0.117029
[v]	-68.05	0.996893
[a]	-69.4	0.954153
[n]	-85.32	0.061519
[i]	-80.85	0.362574
[T]	-89.19	0.04078

Table 4.10: The Confusion Matrix for the Read Mode. The Table Shows The Confusion of Phone Decoding Among 12 languages in Read Mode

	Assamese	Bengali	Gujarati	Hindi	Kannada	Mala	Manipuri	Marathi	Odia	Punjabi	Telugu	Urdu	Training Duration	Testing Duration
Assamese	99.27	0.12	0	0	0	0.36	0	0.24	0	0	0	0	02:00:00	00:30:00
Bengali	0	74.49	14.86	0	0.16	7.93	0	2.19	0.16	0	0.16	0	03:49:55	01:10:55
Gujarati	0	0	98.49	0	0.32	0.32	0	0.53	0.1	0	0.1	0.1	04:09:33	01:24:55
Hindi	0	0.5	0	99.24	0	0.25	0	0	0	0	0	0	02:09:17	00:44:54
Kannada	0	0	32.72	0	67.27	0	0	0	0	0	0	0	01:56:06	00.29.40
Mala	0	12.45	6.38	0.03	0.18	71.09	0.03	3.79	5.81	0	0	0.18	03:58:50	01:13:57
Manipuri	0	0	12.87	0	0	0	86.11	1.01	0	0	0	0	03:28:26	01:31:59
Marathi	0.05	0.05	38.7	0.05	15.93	3.79	0	39.34	0.05	0.05	0	1.92	05:25:41	02:11:54
Odia	0	0	57.71	0	0	0	0	0.93	41.35	0	0	0	03:46:37	01:19:24
Punjabi	0	0	14.12	0	0	0	0	0.47	0.15	85.24	0	0	02:18:12	00:47:44
Telugu	0	0	0.06	0	0	0.33	0	0	0.13	0	77.4	22.05	03:27:19	01:11:12
Urdu	0	0.12	0.6	0.18	0	0.3	0	0.66	0.24	0	13.24	84.64	03:34:50	01:12:13

Table 4.11: The Confusion Matrix for the Lecture Mode. The Table Shows The Confusion of Phone Decoding Among 12 languages in Lecture Mode

	Assamese	Bengali	Gujarati	Hindi	Kannada	Mala	Manipuri	Marathi	Odia	Punjabi	Telugu	Urdu	Training Duration	Testing Duration
Assamese	99.66	0	0	0	0	0	0	0	0.33	0	0	0	02:00:00	00:30:00
Bengali	0	98.61	0.27	0	0	1.1	0	0	0	0	0	0	01:59:29	00:31:14
Gujarati	0	0.25	58.64	0.25	3	2	0.5	0.5	24.56	0	6.76	3.5	02:16:00	00:40:26
Hindi	0	4.76	0	81.74	0	0	0	7.14	0	0	6.34	0	00:43:28	00:17:15
Kannada	0	0	0.25	0	98.45	1.16	0	0	0.12	0	0	0	01:56:11	00.39.14
Mala	0.61	11.86	28.69	2.99	10.7	33.53	0.4	0	6.47	0	4.15	0.54	01:56:11	00:35:41
Manipuri	0	34.23	0	1.8	0	2.7	61.26	0	0	0	0	0	01:39:44	00:50:53
Marathi	0.27	0.27	3.85	0.27	30.02	4.95	0	60.33	0	0	0	0	02:06:51	00:42:16
Odia	0	0	3.79	0	0	1.89	0	0	91.46	0	0.94	1.89	01:51:21	00:37:29
Punjabi	0	2.59	15.58	0	0	6.49	0	0	2.59	71.42	1.29	0	00:15:18	00:04:45
Telugu	0.73	0	0	0.29	0	0	0	0.14	1.75	0	69.06	28	01:29:45	00:30:59
Urdu	0	0	0.51	0	11.73	0	0	0	0	0	18.49	69.26	01:35:39	00:36:08

Table 4.12: The Confusion Matrix for the Conversation Mode. The Table Shows The Confusion of Phone Decoding Among 12 languages in Conversation Mode

	Assamese	Bengali	Gujarati	Hindi	Kannada	Mala	Manipuri	Marathi	Odia	Punjabi	Telugu	Urdu	Training Duration	Testing Duration
Assamese	99.45	0	0	0	0	0	0	0.55	0	0	0	0	02:00:00	00:30:00
Bengali	0	50.12	13.02	19.41	7.61	3.19	0	3.19	0	0	2.94	0.49	02:13:07	00:20:41
Gujarati	0	0	89.37	0.92	1.59	4.24	0.39	0.26	2.52	0.13	0.39	0.13	03:12:07	01:05:35
Hindi	0	0	2.66	69.2	0	0	0	0	0	0	11.02	17.11	02:00:09	00:32:44
Kannada	0	0	0.51	0	94.58	4.63	0.25	0	0	0	0	0	01:52:59	00.27.43
Mala	0	0	39.86	0	13.97	44.25	0.16	0	0	0	1.65	0.08	01:52:59	00:37:19
Manipuri	2.7	0	10.81	0	0	8.1	78.37	0	0	0	0	0	02:03:56	00:27:21
Marathi	0.4	2.44	7.95	1.02	5.51	5.91	0	75.91	0	0	0.61	0.2	01:41:57	00:35:31
Odia	0	0	0	0	0	0	0	0	100	0	0	0	02:13:07	00:20:41
Punjabi	0	0	0	0	26.58	68.35	0	0	0	5.06	0	0	00:15:06	00:05:07
Telugu	0	0	0.1	2.62	0.1	0.8	0	0	1.81	0	94.14	0.4	01:46:51	00:39:01
Urdu	0	0	0.29	19.76	0	0	0	0.29	1.48	0	14.41	63.74	01:42:49	00:31:06

Table 1 : Phonetic Engine Results for 12 languages. The table gives the Information about the Language, Number of Speakers used in Training and Testing, Duration of Training and Testing Data and Phone Recognition Accuracy. The performance of Phonetic Engine for Different Speaking Modes is also given in The Table.

Institute	Language	Train Mode	Test Mode	Training Duration	Testing Duration	Train Speakers	Test Speakers	Accuracy
NEHU	Manipuri	Read	Read	03:28:26	01:31:59	9 (3M+6F)	6 (2M+4F)	62.11%
NEHU	Manipuri	Read	Lect	03:28:26	00:50:53	9 (3M+6F)	2 (1M+1F)	52.27%
NEHU	Manipuri	Read	Conv	03:28:26	00:27:21	9 (3M+6F)	2 (2M)	43.50%
NEHU	Manipuri	Lect	Lect	01:39:44	00:50:53	4 (2M+2F)	2 (1M+1F)	59.78%
NEHU	Manipuri	Lect	Read	01:39:44	00:33:34	4 (2M+2F)	4 (4F)	55.52%
NEHU	Manipuri	Lect	Conv	01:39:44	00:27:21	4 (2M+2F)	2 (2M)	48.39%
NEHU	Manipuri	Conv	Conv	02:03:56	00:27:21	2 (2M)	2 (2M)	53.92%
NEHU	Manipuri	Conv	Read	02:03:56	00:33:34	2 (2M)	4 (4F)	48.41%
NEHU	Manipuri	Conv	Lect	02:03:56	00:50:53	2 (2M)	2 (1M+1F)	47.64%
NEHU	Manipuri	Combine	Read	07:12:06	01:31:59	15 (7M+8F)	6 (2M+4F)	60.53%
NEHU	Manipuri	Combine	Lect	07:12:06	00:50:53	15 (7M+8F)	2 (1M+1F)	61.29%
NEHU	Manipuri	Combine	Conv	07:12:06	00:27:21	15 (7M+8F)	2 (2M)	52.90%
NEHU	Manipuri	Combine	Combine	07:12:06	02:50:13	15 (7M+8F)	10(5M+5F)	59.03%
IITG	Assamese	Read	Read	06:16:20	02:31:20	14 (9M+5F)	6 (6F)	49.60%
IITG	Assamese	Read	Lect	06:35:58	00:52:23	14 (3M+11F)	2 (2F)	43.27%
IITG	Assamese	Read	Conv	06:16:20	00:31:17	14 (3M+11F)	4 (4M)	29.25%
IITG	Assamese	Lect	Lect	03:34:00	00:52:23	4 (1M+3F)	2 (2F)	52.26%
IITG	Assamese	Lect	Read	03:34:00	01:10:58	4 (1M+3F)	3 (3F)	49.96%
IITG	Assamese	Lect	Conv	03:34:00	00:31:17	4 (1M+3F)	4 (4M)	25.66%
IITG	Assamese	Conv	Conv	02:01:39	00:31:17	8 (3M+5F)	4 (4M)	29.82%
IITG	Assamese	Conv	Read	02:01:39	01:20:30	8 (3M+5F)	3 (3F)	47.54%
IITG	Assamese	Conv	Lect	02:01:39	00:52:23	8 (3M+5F)	2 (2F)	42.06%
IITG	Assamese	Combine	Read	09:51:59	02:31:20	26 (13M+13F)	6 (6F)	54.26%
IITG	Assamese	Combine	Lect	09:51:59	00:52:23	26 (13M+13F)	2 (2F)	50.02%
IITG	Assamese	Combine	Conv	09:51:59	00:31:17	26 (13M+13F)	4 (4M)	26.69%
IITG	Assamese	Combine	Combine	09:51:59	03:34:00	26 (13M+13F)	12 (6M+6F)	29.82%
IITKGP	Odia	Read	Read	03:46:37	01:19:24	22 (10M+12F)	9 (4M+5F)	59.69%
IITKGP	Odia	Read	Lect	03:46:37	00:37:29	22 (10M+12F)	2 (1M+1F)	41.85%
IITKGP	Odia	Read	Conv	03:46:37	00:20:41	22 (10M+12F)	2 (1M+1F)	35.03%
IITKGP	Odia	Lect	Lect	01:51:21	00:37:29	6 (3M+3F)	2 (1M+1F)	43.43%
IITKGP	Odia	Lect	Read	01:51:21	01:19:24	6 (3M+3F)	9 (4M+5F)	39.29%
IITKGP	Odia	Lect	Conv	01:51:21	00:20:41	6 (3M+3F)	2 (1M+1F)	33.89%
IITKGP	Odia	Conv	Conv	02:13:07	00:20:41	6 (3M+3F)	2 (1M+1F)	49.58%
IITKGP	Odia	Conv	Read	02:13:07	01:19:24	6 (3M+3F)	9 (4M+5F)	49.98%
IITKGP	Odia	Conv	Lect	02:13:07	00:37:29	6 (3M+3F)	2 (1M+1F)	42.52%
IITKGP	Odia	Combine	Read	07:51:05	01:19:24	34(16M+18F)	9 (4M+5F)	67.67%
IITKGP	Odia	Combine	Lect	07:51:05	00:37:29	34(16M+18F)	2 (1M+1F)	53.59%
IITKGP	Odia	Combine	Conv	07:51:05	00:20:41	34(16M+18F)	2 (1M+1F)	55.26%
IITKGP	Odia	Combine	Combine	07:51:05	02:17:34	34(16M+18F)	13 (6M+7F)	61.71%
IITKGP	Bengali	Read	Lect	03:49:55	00:31:14	38 (14M+24F)	4 (2M+2F)	27.78%
IITKGP	Bengali	Read	Conv	03:49:55	00:38:31	38 (14M+24F)	6 (3M+3F)	16.50%
IITKGP	Bengali	Lect	Lect	01:59:29	00:31:14	7 (5M+2F)	4 (2M+2F)	35.80%
IITKGP	Bengali	Lect	Read	01:59:29	01:10:55	7 (5M+2F)	12 (6M+6F)	36.20%
IITKGP	Bengali	Lect	Conv	01:59:29	00:38:31	7 (5M+2F)	6 (3M+3F)	20.75%
IITKGP	Bengali	Conv	Conv	01:59:11	00:38:31	24 (19M+5F)	6 (3M+3F)	23.99%
IITKGP	Bengali	Conv	Read	01:59:11	01:10:55	24 (19M+5F)	12 (6M+6F)	27.23%
IITKGP	Bengali	Conv	Lect	01:59:11	00:31:14	24 (19M+5F)	4 (2M+2F)	20.89%
IITKGP	Bengali	Combine	Read	06:48:35	01:10:55	69 (38M+31F)	12 (6M+6F)	51.72%
IITKGP	Bengali	Combine	Lect	06:48:35	00:31:14	69 (38M+31F)	4 (2M+2F)	43.24%
IITKGP	Bengali	Combine	Conv	06:48:35	00:38:31	69 (38M+31F)	6 (3M+3F)	30.03%
IITKGP	Bengali	Combine	Combine	06:48:35	02:20:40	69 (38M+31F)	22(11M+11F)	41.87%

Table 1 Continued

Institute	Language	Train Mode	Test Mode	Training Duration	Testing Duration	Train Speakers	Test Speakers	Accuracy
IITH	Telugu	Read	Read	03:27:19	01:11:12	11 (5M+6F)	7 (3M+4F)	56.08%
IITH	Telugu	Read	Lect	03:27:19	00:30:59	11 (5M+6F)	2 (1M+1F)	32.32%
IITH	Telugu	Read	Conv	03:27:19	00:39:01	11 (5M+6F)	13 (10M+3F)	36.52%
IITH	Telugu	Lect	Lect	01:29:45	00:30:59	7 (6M+1F)	2 (1M+1F)	40.20%
IITH	Telugu	Lect	Read	01:29:45	01:11:12	7 (6M+1F)	7 (3M+4F)	43.73%
IITH	Telugu	Lect	Conv	01:29:45	00:39:01	7 (6M+1F)	13 (10M+3F)	37.57%
IITH	Telugu	Conv	Conv	01:46:51	00:39:01	34(30M+4F)	13 (10M+3F)	45.83%
IITH	Telugu	Conv	Read	01:46:51	01:11:12	34(30M+4F)	7 (3M+4F)	49.07%
IITH	Telugu	Conv	Lect	01:46:51	00:30:59	34(30M+4F)	2 (1M+1F)	32.13%
IITH	Telugu	Combine	Read	06:43:55	01:11:12	52(41M+11F)	7 (3M+4F)	55.38%
IITH	Telugu	Combine	Lect	06:43:55	00:30:59	52(41M+11F)	2 (1M+1F)	37.66%
IITH	Telugu	Combine	Conv	06:43:55	00:39:01	52(41M+11F)	13 (10M+3F)	42.27%
IITH	Telugu	Combine	Combine	06:43:55	02:21:12	52(41M+11F)	22(14M+8F)	48.35%
IITH	Urdu	Read	Read	03:34:50	01:12:13	32 (29M+3F)	25 (22M+3F)	45.67%
IITH	Urdu	Read	Lect	03:34:50	00:36:08	32 (29M+3F)	1 (1M)	50.28%
IITH	Urdu	Read	Conv	03:34:50	00:31:06	32 (29M+3F)	7 (4M+3F)	46.16%
IITH	Urdu	Lect	Lect	01:35:39	00:36:08	3 (2M+1F)	1 (1M)	54.83%
IITH	Urdu	Lect	Read	01:35:39	01:12:13	3 (2M+1F)	25 (22M+3F)	39.16%
IITH	Urdu	Lect	Conv	01:35:39	00:31:06	3 (2M+1F)	7 (4M+3F)	34.49%
IITH	Urdu	Conv	Conv	01:42:49	00:31:06	16 (8M+8F)	7 (4M+3F)	48.84%
IITH	Urdu	Conv	Read	01:42:49	01:12:13	16 (8M+8F)	25 (22M+3F)	41.17%
IITH	Urdu	Conv	Lect	01:42:49	00:36:08	16 (8M+8F)	1 (1M)	49.69%
IITH	Urdu	Combine	Read	06:51:18	01:12:13	54 (39M+12F)	25 (22M+3F)	46.02%
IITH	Urdu	Combine	Lect	06:51:18	00:36:08	54 (39M+12F)	1 (1M)	54.74%
IITH	Urdu	Combine	Conv	06:51:18	00:31:06	54 (39M+12F)	7 (4M+3F)	48.32%
IITH	Urdu	Combine	Combine	06:51:18	02:19:27	54 (39M+12F)	33 (27M+7F)	48.55%
TEZU	Assamese	Read	Read	01:24:56	00:10:18	11 (3F+8M)	2 (2F)	51.12%
TEZU	Assamese	Read	Lect	01:24:56	00:11:48	11 (3F+8M)	2 (2F)	44.12%
TEZU	Assamese	Read	Conv	01:24:56	00:06:34	11 (3F+8M)	2 (2M)	36.43%
TEZU	Assamese	Lect	Lect	01:46:48	00:11:48	12 (3F+9M)	2 (2F)	43.87%
TEZU	Assamese	Lect	Read	01:46:48	00:10:18	12 (3F+9M)	2 (2F)	47.98%
TEZU	Assamese	Lect	Conv	01:46:48	00:06:34	12 (3F+9M)	2 (2M)	39.65%
TEZU	Assamese	Conv	Conv	00:16:33	00:06:34	6 (6M)	2 (2M)	35.01%
TEZU	Assamese	Conv	Read	00:16:33	00:10:18	6 (6M)	2 (2F)	33.48%
TEZU	Assamese	Conv	Lect	00:16:33	00:11:48	6 (6M)	2 (2F)	35.01%
TEZU	Assamese	Combine	Read	03:28:17	00:10:18	29 (23M+06F)	2 (2F)	50.23%
TEZU	Assamese	Combine	Lect	03:28:17	00:11:48	29 (23M+06F)	2 (2F)	44.83%
TEZU	Assamese	Combine	Conv	03:28:17	00:06:24	29 (23M+06F)	2 (2M)	38.66%
TEZU	Assamese	Combine	Combine	03:28:17	00:28:00	29 (23M+06F)	6 (2M+4F)	50.04%
DAIICT	Marathi	Read	Read	05:25:41	02:11:54	61 (42M+19F)	23 (18M+5F)	52.82%
DAIICT	Marathi	Read	Lect	05:25:41	00:42:16	61 (42M+19F)	2 (1M+1F)	33.49%
DAIICT	Marathi	Read	Conv	05:25:41	00:35:31	61 (42M+19F)	10 (4M+6F)	36.87%
DAIICT	Marathi	Lect	Lect	02:06:51	00:42:16	7 (3M+4F)	2 (1M+1F)	35.18%
DAIICT	Marathi	Lect	Read	02:06:51	02:11:54	7 (3M+4F)	23 (18M+5F)	32.97%
DAIICT	Marathi	Lect	Conv	02:06:51	00:35:31	7 (3M+4F)	10 (4M+6F)	27.48%
DAIICT	Marathi	Conv	Conv	01:41:57	00:35:31	26 (19M+7F)	10 (4M+6F)	35.61%
DAIICT	Marathi	Conv	Read	01:41:57	02:11:54	26 (19M+7F)	23 (18M+5F)	45.95%
DAIICT	Marathi	Conv	Lect	01:41:57	00:42:16	26 (19M+7F)	2 (1M+1F)	38.03%
DAIICT	Marathi	Combine	Read	09:14:29	02:11:54	94(64M+30F)	23 (18M+5F)	52.67%
DAIICT	Marathi	Combine	Lect	09:14:29	00:42:16	94(64M+30F)	2 (1M+1F)	36.25%
DAIICT	Marathi	Combine	Conv	09:14:29	00:35:31	94(64M+30F)	10 (4M+6F)	35.95%
DAIICT	Marathi	Combine	Combine	09:14:29	03:29:41	94(64M+30F)	35(23M+12F)	38.84%

Table 1 Continued

Institute	Language	Train Mode	Test Mode	Train Duration (Hrs:Mins:Secs)	Test Duration (Hrs:Mins:Secs)	Train Speakers	Test Speakers	Accuracy
DAIICT	Gujarati	Read	Read	04:09:33	01:24:55	95 (54M+41F)	28 (17M+11F)	63.55%
DAIICT	Gujarati	Read	Lect	04:09:33	00:40:26	95 (54M+41F)	4 (3M+1F)	50.40%
DAIICT	Gujarati	Read	Conv	04:09:33	01:05:35	95 (54M+41F)	46 (26M+20F)	54.58%
DAIICT	Gujarati	Lect	Lect	02:16:00	00:40:26	5 (4M+1F)	4 (3M+1F)	51.52%
DAIICT	Gujarati	Lect	Read	02:16:00	01:24:55	5 (4M+1F)	28 (17M+11F)	51.91%
DAIICT	Gujarati	Lect	Conv	02:16:00	01:05:35	5 (4M+1F)	46 (26M+20F)	45.39%
DAIICT	Gujarati	Conv	Conv	03:12:07	01:05:35	72 (37M+35F)	46 (26M+20F)	55.35%
DAIICT	Gujarati	Conv	Read	03:12:07	01:24:55	72 (37M+35F)	28 (17M+11F)	60.00%
DAIICT	Gujarati	Conv	Lect	03:12:07	00:40:26	72 (37M+35F)	4 (3M+1F)	52.27%
DAIICT	Gujarati	Combine	Read	09:37:40	01:24:55	172(95M+77F)	28 (17M+11F)	62.24%
DAIICT	Gujarati	Combine	Lect	09:37:40	00:40:26	172(95M+77F	4 (3M+1F)	54.69%
DAIICT	Gujarati	Combine	Conv	09:37:40	01:05:35	172(95M+77F	46 (26M+20F)	43.67%
DAIICT	Gujarati	Combine	Combine	09:37:40	03:10:56	172(95M+77F	78(46M+32F)	58.29%
Thapar	Pujabi	Read	Lect	02:18:12	00:04:45	3 (1M+2F)	1 (1M)	26.25%
Thapar	Pujabi	Read	Conv	02:18:12	00:05:07	3 (1M+2F)	5 (4M+1F)	19.46%
Thapar	Pujabi	Lect	Lect	00:15:18	00:04:45	1 (1M)	1 (1M)	47.17%
Thapar	Pujabi	Lect	Read	00:15:18	00:47:44	1(1M)	2 (2M)	20.75%
Thapar	Pujabi	Lect	Conv	00:15:18	00:05:07	1 (1M)	5 (4M+1F)	20.33%
Thapar	Pujabi	Conv	Conv	00:15:06	00:05:07	10 (9M+1F)	5 (4M+1F)	22.39%
Thapar	Pujabi	Conv	Read	00:15:06	00:47:44	10 (9M+1F)	2 (2M)	20.33%
Thapar	Pujabi	Conv	Lect	00:15:06	00:04:45	10 (9M+1F)	1 (1M)	18.53%
Thapar	Pujabi	Combine	Read	02:48:36	00:47:44	14 (11M+4F)	2 (2M)	61.59%
Thapar	Pujabi	Combine	Lect	02:48:36	00:04:45	14 (11M+4F)	1 (1M)	45.06%
Thapar	Pujabi	Combine	Conv	02:48:36	00:05:07	14 (11M+4F)	5 (4M+1F)	22.24%
Thapar	Pujabi	Combine	Combine	02:48:36	00:57:36	14 (11M+4F)	8 (7M+1F)	55.23%
RITK	Malayalam	Read	Read	03:58:50	01:13:57	22 (15M+7F)	8 (3M+5F)	33.23%
RITK	Malayalam	Read	Lect	03:58:50	00:35:41	22 (15M+7F)	6 (3M+3F)	31.70%
RITK	Malayalam	Read	Conv	03:58:50	00:37:19	22 (15M+7F)	8 (3M+5F)	28.77%
RITK	Malayalam	Lect	Lect	01:56:11	00:35:41	13 (10M+3F)	6 (3M+3F)	31.82%
RITK	Malayalam	Lect	Read	01:56:11	01:13:57	13 (10M+3F)	8 (3M+5F)	27.33%
RITK	Malayalam	Lect	Conv	01:56:11	00:37:19	13 (10M+3F)	8 (3M+5F)	32.27%
RITK	Malayalam	Conv	Conv	01:52:59	00:37:19	41 (31M+10F)	8 (3M+5F)	31.18%
RITK	Malayalam	Conv	Read	01:52:59	01:13:57	41 (31M+10F)	8 (3M+5F)	26.31%
RITK	Malayalam	Conv	Lect	01:52:59	00:35:41	41 (31M+10F)	6 (3M+3F)	31.55%
RITK	Malayalam	Combine	Read	03:58:50	01:13:57	76 (56M+20F)	8 (3M+5F)	32.28%
RITK	Malayalam	Combine	Lect	01:56:11	00:35:41	76 (56M+20F)	6 (3M+3F)	33.20%
RITK	Malayalam	Combine	Conv	01:52:59	00:37:19	76 (56M+20F)	8 (3M+5F)	31.98%
RITK	Malayalam	Combine	Combine	07:46:00	02:25:57	76 (56M+20F)	22(9M+13F)	32.47%
SITT	Kannada	Read	Read	01:56:06	00.29.40	11 (5M+6F)	2 (1M+1F)	55.10%
SITT	Kannada	Read	Lect	01:56:06	00.39.14	11 (5M+6F)	2 (1M+1F)	34.77%
SITT	Kannada	Read	Conv	01:56:06	00.27.43	11 (5M+6F)	6 (6M)	40.00%
SITT	Kannada	Lect	Lect	01:56:11	00.39.14	4 (3M+1F)	2 (1M+1F)	48.74%
SITT	Kannada	Lect	Read	01:56:11	00.29.40	4 (3M+1F)	2 (1M+1F)	46.28%
SITT	Kannada	Lect	Conv	01:56:11	00.27.43	4 (3M+1F)	6 (6M)	37.87%
SITT	Kannada	Conv	Conv	01:52:59	00.27.43	12 (11M+1F)	6 (6M)	45.67%
SITT	Kannada	Conv	Read	01:52:59	00.29.40	12 (11M+1F)	2 (1M+1F)	42.99%
SITT	Kannada	Conv	Lect	01:52:59	00.39.14	12 (11M+1F)	2 (1M+1F)	42.29%
SITT	Kannada	Combine	Read	05:44:16	00.29.40	11 (5M+6F)	2 (1M+1F)	53.26%
SITT	Kannada	Combine	Lect	05:44:16	00.39.14	4 (3M+1F)	2 (1M+1F)	46.60%
SITT	Kannada	Combine	Conv	05:44:16	00.27.43	12 (11M+1F)	6 (6M)	44.15%
SITT	Kannada	Combine	Combine	05:44:16	01:36:37	27(19M+8F)	10(8M+2F)	44.15%

Table 1 Continued

Institute	Language	Train Mode	Test Mode	Train Duration (Hrs:Mins:Secs)	Test Duration (Hrs:Mins:Secs)	Train Speakers	Test Speakers	Accuracy
IITK	Hindi	Read	Read	02:09:17	00:44:54	29 (7M+22F)	14 (8M+6F)	49.24%
IITK	Hindi	Read	Lect	02:09:17	00:17:15	29 (7M+22F)	8 (8F)	28.98%
IITK	Hindi	Read	Conv	02:09:17	00:32:44	29 (7M+22F)	9 (5M+4F)	34.06%
IITK	Hindi	Lect	Lect	00:43:28	00:17:15	9 (8M+1F)	8 (8F)	37.96%
IITK	Hindi	Lect	Read	00:43:28	00:44:54	9 (8M+1F)	14 (8M+6F)	33.27%
IITK	Hindi	Lect	Conv	00:43:28	00:32:44	9 (8M+1F)	9 (5M+4F)	39.86%
IITK	Hindi	Conv	Conv	02:00:09	00:32:44	18 (9M+9F)	9 (5M+4F)	42.15%
IITK	Hindi	Conv	Read	02:00:09	00:44:54	18 (9M+9F)	14 (8M+6F)	30.00%
IITK	Hindi	Conv	Lect	02:00:09	00:17:15	18 (9M+9F)	8 (8F)	33.78%
IITK	Hindi	Combine	Read	04:52:54	00:44:54	56 (24M+32)	14 (8M+6F)	46.54%
IITK	Hindi	Combine	Lect	04:52:54	00:17:15	56 (24M+32)	8 (8F)	36.44%
IITK	Hindi	Combine	Conv	04:52:54	00:32:44	56 (24M+32)	9 (5M+4F)	41.06%
IITK	Hindi	Combine	Combine	04:52:54	01:34:53	56 (24M+32)	31 (13M+18M)	43.01%

5

Tezpur University

PROGRESS SUMMARY REPORT OF TEZPUR UNIVERSITY

A. General

A.1	Name of the Project	:	**Development of Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages (Assamese)**
	Our Reference Letter No	:	11(6)/2011-HCC(TDIL) dated 23-12-2011
A.2	Executing Agency	:	Tezpur University
A.3	Chief Investigator with	:	Dr. Utpal Sharma
	Designation	:	Professor
	Co-Chief Investigators with	:	Dr. Smriti Kumar Sinha
	Designation	:	Professor
A.4	Project staffs with Qualification	:	Mr. Navanath Saharia (MSc (CS) and pursuing PhD), Bhaskarjyoti Das (B Tech (CSE)), Nirman Singh (MTech (IT)), Himangshu Sarma(M Tech (IT)), Mancha Jyoti Malakar(MCA), Ms Sanghamitra Nath (M Tech(ECE), Asstt. Prof)
A.5	Total Cost of the Project as approved by DIT	:	
	i) Original	:	30.015 Lakhs
	ii) Revised, if any	:	
A.6	Date of starting (Indicate date of first sanction)	:	23-12-2011
A.7	Date of Completion	:	
	i) Original	:	22-12-2013
	ii) Revised, if any	:	31-03-2015
A.8	Date on which last progress report was Submitted	:	07-03-2014

B. Technical

B.1 Works Progress.(given details in technical report) :

- Database collection in three different modes:
  i. Read speech : 13 hours 27 minutes
     * AIR News : 11 hours 27 minutes
     * Recorded : 2 hours
  ii. Lecture mode : 3 hours 10 minutes
  iii. Conversational speech : 5 hour
- Transcription using IPA chart : 12 hours
- Development of prosody knowledge : We have studied the prosody events in Assamese speech and carryied out pitch marking and break marking using the framework discussed and finalized at IIIT-H.
- Development of phonetic engine : Developed a preliminary phonetic engine based on Hidden Markov Model.
- Development of speech search application: We could develop a simple speech search application that performs speech search after phonetic transcription.

C. Financial

Statement of expenditure (in lakhs)
(Utilization certificate enclosed)

Sl. No	**Sanctioned Heads**	**Funds Received**	**Expenditure incurred (23.12.11 to 30.09.12)**	**Balance**
		A	**B**	**A−B**
1.	Capital Equipment	5.00	4.95	0.05
2.	Data Collection	5.50	5.42	0.08
3.	Consumable stores	2.60	2.60	0.00
4.	Manpower	6.00	6.00	0.00
5.	Travel	2.00	1.98	0.02
6.	Workshop & Training	2.00	2.00	0.00
7.	Contingency	3.00	3.23	-0.23
8.	Coordination & Management	0.00	0.00	0.00
9.	System Integration	0.00	0.00	0.00
10.	Sub Total	26.10	26.18	-0.08
11.	Over Head (15%)	3.915	3.915	0.00
12.	**Total Budget**	**30.015**	**30.09**	**-0.08**

D. Project Outcomes

Development of Database

- **Transcribed Database**: 12 hours.
    i. Read speech : 6 hours
        * AIR News : 4 hours 15 minutes
        * Recorded : 1 hours 39 minutes
    ii. Lecture mode : 3 hours 40 minutes
    iii. Conversational speech : 2 hour 30 minutes

Tools & Systems Developed

- **International Phonetics Alphabet Typing Tool** : An offline Unicode compatible International Phonetics Alphabet (IPA) typing tool useful for phonetic transcription.
- **Phonetic engine using Hidden Markov Model** : Using HTK toolkit.
- **Speech Search Engine** : Search query words in speech database in phonetically transcribed form.
- **Automatic syllabifier**: Identify syllables in phonetic transcribed data using simple rules.
- **Automatic Breakmarking**: Identify silence portions longer than a threshold period.

Papers Published:

[1] Sarma Himangshu, Saharia Navanath, Sharma Utpal. "Development of Assamese Speech Corpus and Automatic Transcription Using HTK" "International Symposium on Signal Processing and Intelligent Recognition Systems, Thiruvananthapuram, India, 2014. Available in *Advances in Intelligent Systems and Computing*, Springer Link, Volume 264, 2014, pp 119-132, March 2014.

[2] Nath Sanghamitra, Sarma Himangshu, Sharma Utpal. "Assamese Dialect Translation System- A preliminary proposal", International Conference on Computer Science, Engineering and Applications (ICONACC)-2014, Manipur University, Imphal, March 10 - 12, 2014.

[3] Nath Sanghamitra, Sarma Himangshu, Sharma Utpal. "A study of VOT patterns in Assamese and its Nalbaria variety", "Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2014)", Nepal, 2014. Available in *Lecture Notes in Computer Science (LNCS)8404*.

[4] Sarma Himangshu, Saharia Navanath, Sharma Utpal, Sinha Smriti Kumar, Malakar Mancha Jyoti. Development and Transcription of Assamese Speech Corpus, National seminar cum Conference on Recent threads and Techniques in Computer Sciences. Bodoland University, India, 2013

Detailed Technical Report of Tezpur University

0.1 Database collection & transcription

Data Collection We recorded over 13 hours read speech by about 25 different speakers. Of this, about 11 hours data was collected from AIR News with 10 different speakers, and around 2 hours read speech was recorded from about 15 native speakers (20-40 age group) using Sony ICD-UX533F recording device.

We also recorded over 3 hour data in lecture mode including about 2 hours of extempore speech, from 5 different speakers with average 15 minutes duration. All are native speakers of Assamese in the age group 20-40.

Further, we recorded over five hours of speech data in conversation mode. Around 1 hours 30 minutes of this is an interview with 2 persons. Another 1 hour interview speech involving four different speakers were collected from two news channels.

All recordings were done in closed room environment and saved with 44.1 KHz sampling rate.

Transcription Phonetic transcription using the International Phonetic Alphabet (IPA) symbols has been done manually for about 12 hours speech data. This includes of 6 hours of read speech data, 3 hours and 40 minutes of extempore speech and 2 hours and 10 minutes of conversation data. The transcription was done by the project staff as well as some students of MA (Linguitistics) programme. The transcription has been cross verified within the group.

The list of IPA symbols used in transcription is shown in Table 1.

During transcription we faced different problems and some interesting observation related to consonants and vowels. The pronunciation rules are different from different regions of Assam. The following are some interesting observations and problems we faced.

- ***Consonants***

  (i) w is used for ৱ if ৱ is used in the middle or the beginning of a word, but b is used for ৱ if ৱ is the last letter of the word. e.g.

  ছোৱালী = ʃowali ঘৰুৱা = g^hɒɹuwa ৱাহিদ = wahid ৱাল = wal দেৱ = debɒ কেশৱ = kexɒbɒ

  (ii) ʧ is not used in proper Assamese pronunciation, but it is used in English, Bengali and Hindi languages. Now a days many Assamese people speak English, Bengali and Hindi. When such people speak Assamese they use ʧ. e.g.
  আচ্চা = aʧa বাচ্চা = baʧa

	Letter	IPA	Letter	IPA	Letter	IPA
	ক	k	ণ	n	ৱ	w / bɒ
	খ	k^h	ত	t	শ	x
	গ	g	থ	t^h	ষ	x
C	ঘ	g^h	দ	d	স	s / x
O	ঙ	ŋ	ধ	d^h	হ	h
N	চ	s / ʃ / ʧ	ন	n	ক্ষ	k^hj
S	ছ	s / ʃ / ʧ	প	p	য়	j
O	জ	ʤ	ফ	p^h	ড়	ɹ
N	ঝ	$ʤ^h$	ব	b	ঢ়	ɹh
A	ঞ	ɲ	ভ	b^h	ৎ	t̪
N	ট	t	ম	m	ং	ŋ
T	ঠ	t^h	য	ʤ	ঃ	'
S	ড	d	ৰ	ɹ	ঁ	~
	ঢ	d^h	ল	l		
V						
O	অ	ɔ / ɒ	ঊ	u	ঐ	oi
W	আ	a	ঊ	u	ও	ʊ / o
E	ই	i	ঋ	ɹi	ঔ	oʊ
L	ঈ	i	এ	e / ɛ		
S						

Table 1: IPA symbols for Assamese letters.

(iii) During transcription we saw that a ɒ occurs after every consonant, but some consonants viz. ঙ, ৎ, ং, ঃ do not have it. t e.g.

অংক = ɒŋkɒ আঙুৰ = ɒŋuɹ উৎসৱ = ut̪xɒb নিঃকিন = nikin

(iv) Another interesting observation is for the letter স. When we use স as a single letter the corresponing IPA symbol is x, but when we use স in a cluster the IPA symbol is s. e.g.
সাধাৰণ = xad^{h}aɹɒn সাবতি = xabɒti ব্যৱস্থাত = bjbɒst^{h}at

(v) The presence of ɒ after a consonant is irregular. In some words ɒ occurs after a consonant and in others it does not. E.g. জ
ৰাইজক = ɹaiʤɒk জগতৰ = ʤɒgɒtɒɹ ৰাজপথ = ɹaʤpɒt^h ৰাজগড় = ɹaʤgɒɹ

- ***Vowels***

  The vowel sounds ʊ and u are sometimes used for same word when they are spoken by people from different regions. E.g.

  বোলেও = bʊleo বুলেও = buleo তোমাৰ = tʊmaɹ তুমাৰ = tumaɹ

- ***Clustering***

  In general, if consonant cluster occurs at the final position of a word the vowel ɒ is required at the last position of the word. e.g.

মন্তব্য = mɒntɒbjɒ অস্তিত্ব = ɒstitbɒ ৰক্ত = ɹɒktɒ

But, in some cases if a consonant cluster occurs at final position of a word ɒ is not required. e.g.

কান্ধ = kand^{h} ৰান্ধ = band^{h}

Tools developed

- ***IPA Tool***: For phonetic transcription we developed a standalone International Phonetics Alphabet Typing Tool in C++. It removes several drawbacks of other existing methods. Particularly the existing online IPA typing tools do not support the new IPA symbols. For the consortium members the tool is available for download from http://www.tezu.ernet.in/∼nlp/ipa.htm.

- ***Phonetic Engine*** We implemented a simple phonetic engine based on Hidden Markov Model, using idea developed in IIT Guwahati. It uses the HTK toolkit. For faster interactive response the input is provided as speech segments of less than 10 seconds. We use 38 basic phones represented by ASCII symbols. The list of ASCII symbols used in phonetic engine is shown in Table 2.

  We obtained transcription accuracy of around 60%.

- ***Speech Search Engine*** We could develop a simple speech search engine that takes a word as a query and locates segments from the speech database that contain that word. It performs the search after phonetic transcription of the query word as well as texts in the search database.

- ***Automatic Syllabifer*** We developed a simple syllabifier for Assamese speech, that identifies syllables in phonetic transcribed data using simple rules.

- ***Automatic Breakmarking*** We developed an automatic break marking application, that identifies in the speech silence portions longer than a threshold period.

0.2 Acquiring prosody knowledge

While carrying out phonetic transcription of Assamese speech we found that same word is spoken differently by speakers in different contexts. Some of these differences need to be transcribed using different IPA symbols, and

others require other distinct prosody labels, to distinctly represent the pronunciation. One such prosody feature we have studied is the Voice Onset Time (VOT).

We have also analyzed different dialects of the Assamese language which gives a unique idea of prosody of the language.

0.3 Development of phonetic engine

In this project we could develop some very essential resources for speech processing research, particularly for working with Assamese speech. For such work a good speech corpus as well as a manually tagged corpus is highly desirable. We have developed an Assamese speech database of about 21 hours with three broad types of speech- read speech, lecture and conversation. We have manually transcribed about 12 hours of the data using IPA symbols. For closer analysis of speech, the prosody is important. We have done prosody labelling of part of the corpus. Particularly, we have done pitch-marking, break-marking and syllabification. In terms of application development, we implemented an HMM based phonetic engine in collaboration with IIT Guwahati, and a speech search engine that essentially works over transcribed speech data. We also developed an IPA typing tool that is useful in phonetic transcription.

With the experience gained in this project, we realise that prosody information can be used to improve speech analysis. An interesting work can be regarding dialectal differences in the spoken form of a language.

0.4 Development of speech search engine

We have developed a speech search engine using our phonetic engine. Firstly we convert the spoken search key to IPA transcribed form using our phonetic engine. Then we compare the string matching with our database where speech files and their respective transcribed files are stored. After comparison, matching .wav files are presented as results. We observe that exact matching is not very effective, and we intend to experiment with approximate matching.

0.5 Summary & Future work

From the recorded and AIR speech data, we have transcribed with prosody marking these using a tool that we have developed and also we have done break-marking, some amount of pitch-marking and syllabification of these data. We have developed a phonetic engine using Hidden Markov Model. The developed phonetic engine give an accuracy of more than 60%.

	Letter	Symbol	Letter	Symbol	Letter	Symbol
	ক	k	ণ	n	ৱ	w
	খ	kh	ত	t	শ	x
C	গ	g	থ	th	ষ	x
O	ঘ	gh	দ	d	স	x
N	ঙ	ng	ধ	dh	হ	h
S	চ	s	ন	n	ক্ষ	khy
O	ছ	s	প	p	য়	y
N	জ	j	ফ	ph	ড়	r
A	ঝ	j	ব	b	ঢ়	rh
N	ঞ	yo	ভ	bh	ৎ	ta
T	ট	t	ম	m	ং	ng
S	ঠ	th	য	j		
	ড	d	ৰ	r		
	ঢ	dh	ল	l		
V						
O	অ	oa	উ	u	ঐ	oi
W	আ	a	ঊ	u	ও	o
E	ই	i	ঋ	ri	ঔ	ou
L	ঈ	i	এ	e		
S						

Table 2: Representation of Assamese sound in ASCII; where L−Letters of Assamese alphabet

We have also tried to develop a speech search engine which give a combination of speech files and IPA transcription as the output. We also try to make some rule for automatic syllabification of Assamese speech which give an accuracy of more than 90% accuracy.

0.6 References

6

North Eastern Hill University (NEHU)

Shillong

PROGRESS SUMMARY REPORT OF NEHU Shillong

A. General

A.1	Name of the Project	:	**Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages (Manipuri)**
	Our Reference Letter No	:	11(6)/2011-HCC(TDIL) dated 23-12-2011
A.2	Executing Agency	:	NEHU Shillong
A.3	Chief Investigator with Designation	:	Dr. L. Joyprakash Singh Associate Professor
	Co-Chief Investigators with Designation	:	Mr. Sushanta Kabir Dutta Associate Professor
A.4	Project staffs with Qualification	:	Mr. Salam Nandakishor, B.E.(EC)
A.5	Total Cost of the Project as approved by DIT	:	
	i) Original	:	Rs. 30.015 Lakhs
	ii) Revised, if any	:	
A.6	Date of starting (Indicate date of first sanction)	:	23/12/2011
A.7	Date of Completion	:	
	i) Original	:	22/12/2013
	ii) Revised, if any	:	
A.8	Date on which last progress report was Submitted	:	28/02/2014

B. Technical

B.1 Works Progress.(given details in technical report) :

- Database collection in three different modes:
  i. Read speech : 10 Hrs
  ii. Lecture mode : 5 Hrs
  iii. Conversational speech : 5 Hrs
- Transcription using IPA chart : 10 hrs of data has been transcribed.
- Development of prosody knowledge : Completed
- Development of phonetic engine : Has been developed

B.2 Proposed plan-of-work highlighting the action to be taken to achieve the originally proposed targets :

- Report finalization
- Database finalization
- Code delivery
- Finance settlement

C. Financial

Sl. No	**Sanctioned Heads**	**Funds Received (in lakhs)**	**Expenditure incurred (23.12.2011 to 22.12.2014) (in lakhs)**	**Balance (in lakhs)**
		A	**B**	**A−B**
1.	Capital Equipment	5.0	5.11722	-0.11722
2.	Data Collection	5.5	5.02645	0.47355
3.	Consumable stores	2.6	2.63317	-0.03317
4.	Manpower	6.0	5.59129	0.40871
5.	Travel	2.0	1.56581	0.43419
6.	Workshop & Training	2.0	1.90236	0.09764
7.	Contingency	3.0	4.26393	-1.26393
8.	Coordination & Management	0.0	-	-
9.	System Integration	0.0	-	-
10.	Sub Total	25.1	26.10023	-0.00023
11.	Over Head (15%)	3.915	3.91500	NIL
12.	**Total Budget**	**30.015**	**30.01523**	**-0.00023**

D. Project Outcomes

Papers Published(Symposium/Conference):

- Salam Nandakishor, Laishram Rahul, S.K Dutta and L. Joyprakash Singh, *"Development of Manipuri Phonetic Engine"*, Zonal Seminar, The Institute of Electronics and Telecommunication Engineers [IETE], May 3-4, 2013.
- Laishram Rahul, Salam Nandakishor, L. Joyprakash Singh and S.K.Dutta, *"Design of Manipuri Keywords Spotting System using HMM"*, NCVPRIPG 2013, December 18-21, 2013 at IIT Jodhpur, IEEE Proceedings, 2013.

Development of Database

- Manipuri speech database has been created. It consists of 3 sections : SPC-1, SPC-2, SPC-3. One hour of speech transcription with prosody marking has been included as a part of SPC-3. SPC-1, SPC-2 and SPC-3 contain 3, 20 and 10 hours of speech database in Read, Lecture and Conversation mode respectively. Meta data for each wave file has also been created. Two groups of words of about 30 and 376 words have selected as query and vocabulary words respectively have been made available in two separate files namely query.txt and vocabulary.txt.

Tools & Systems Developed

- Phonetic engine for Manipuri language is developed using HTK v3.4.
- Manipuri Keywords Spotting System has been developed.
- An HMM based Semi-Automatic syllable labeling system for Manipuri language has been developed.
- GUI has been designed for Prosodically guided Manipuri Phonetic Engine.

Appendix 6.1

Detailed Technical Report of NEHU Shillong

Detailed Technical Report of NEHU Shillong

6.1 Database collection & transcription

1. Data Collection: A good quality data of about 10 hours in read speech has been collected from the recording studio as well as from the AIR imphal. This data consists of speech read by male and female speakers. During collection of data, care has been taken to maintain wherever possible an equal amount of data from both male and female speakers. The H4n recording devices have been used during recording in the studio. The device is maintained at a sampling frequency of 48 kHz and 44.1 kHZ, 16 bit per sample size and WAV format. We have also collected about 5 hrs of data in conversational mode of speech and another group of data of about 5 hrs in lecture mode. These data have been collected from fields and the studio.

2. Transcription: The broadcast data acquired has been chunked into smaller parts proportionate to the length of a sentence. Each chunked data are listened and analyzed carefully to obtain higher accuracy in transcription. Transcription of 5, 2.5 and 2.5 hours of data has been done on Read, Conversation, Lecture mode data respectively. The Read mode data has been collected from 4 males and 10 females native speakers of Manipuri language. Each of these male speakers used about 37 phones while among female speakers, three used 37 phones each while others used 36 phones respectively. International Phonetic Alphabet (IPA) chart (2005 revision) is being used during transcription process. A total of 38 phones have been used by speakers altogether. In Conversation mode, data are recorded from 7 male speakers and 37 phones have been used. The Lecture mode data are collected from 2 female speakers and speakers used 37 phones.

6.2 Acquiring prosody knowledge

Pitch marking, prosodic Break marking and Syllabification have been done on speech data of Read, Conversation and Lecture mode respectively. To show prosody information on speech data, a Graphical User Interface (GUI) has been designed for Prosodically Guided Manipuri Phonetic Engine. The snapshot of the main interface of the GUI is shown in Figure-6.1 while prosody outputs are shown in Figure-6.2 - Figure-6.4.

Figure 6.1: Main interface of the GUI.

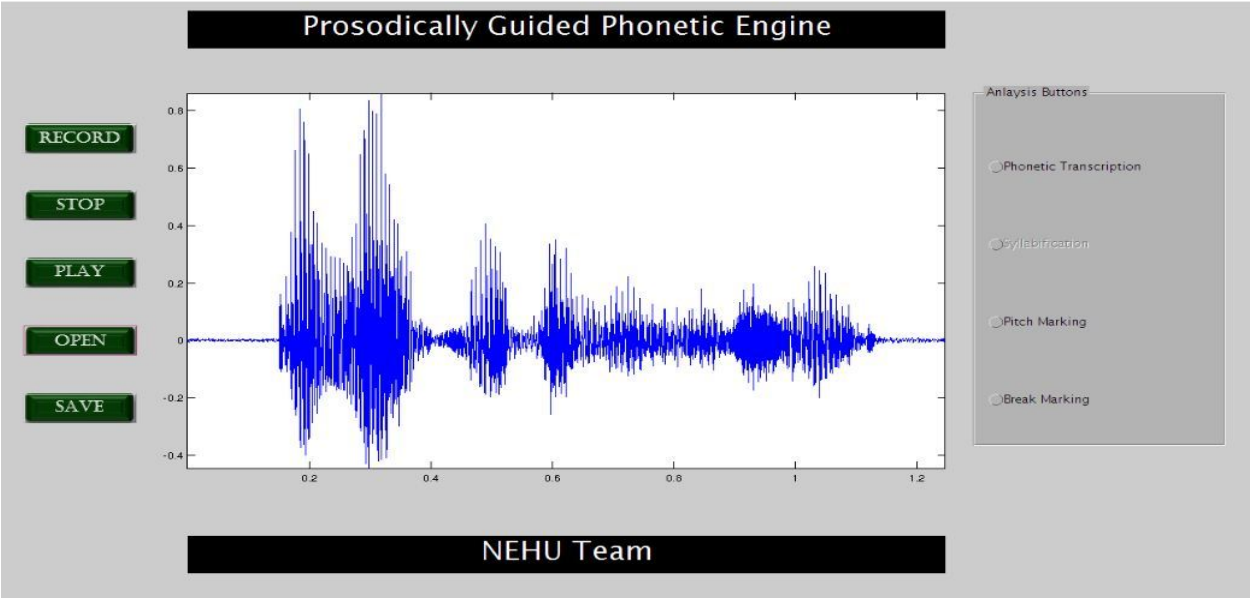


Figure 6.2: Phonetic Transcription GUI output

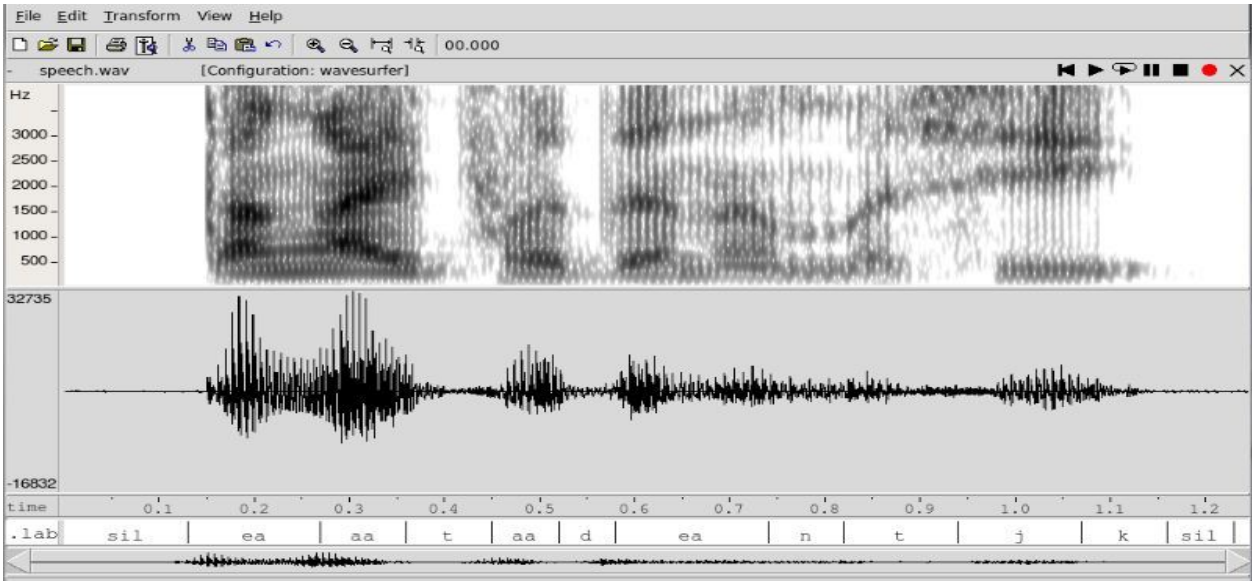


Figure 6.3: Pitch Marking GUI output

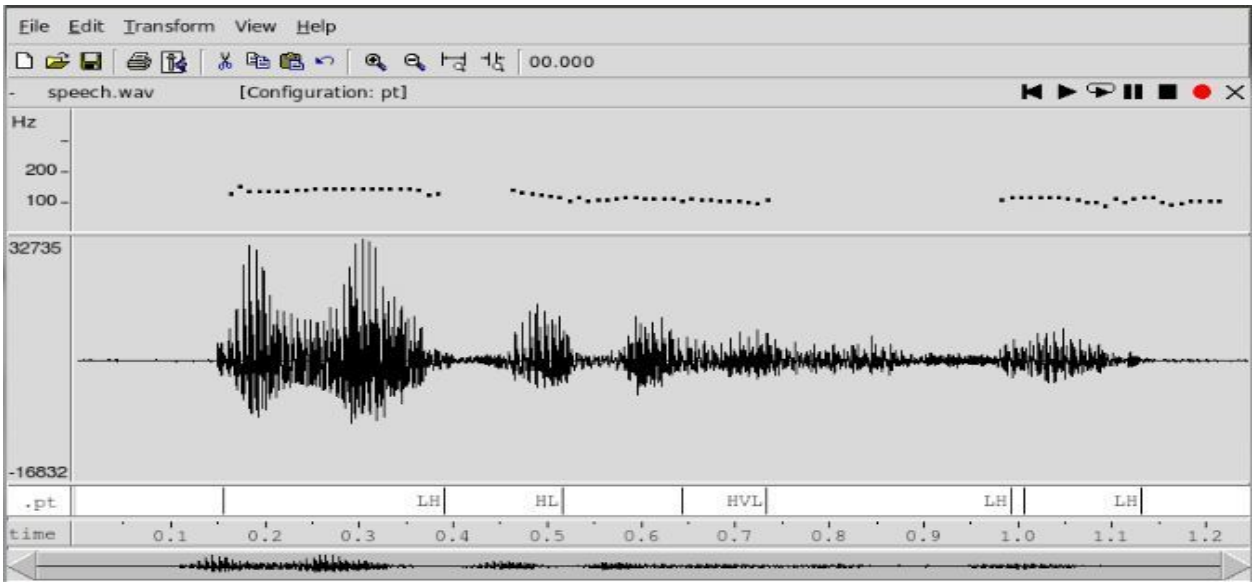
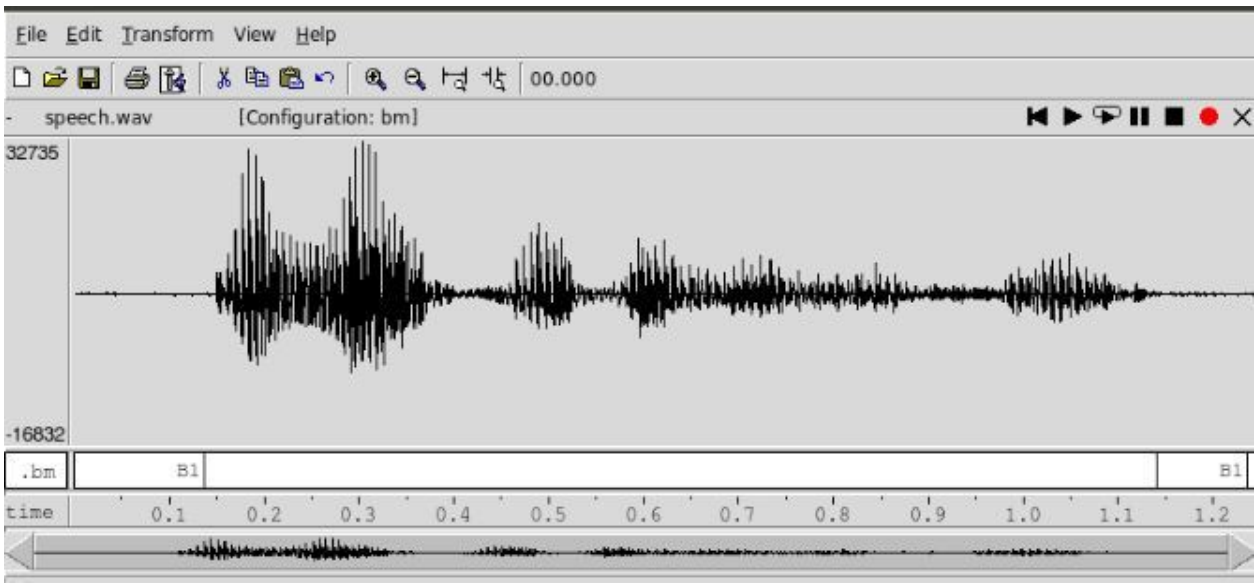


Figure 6.4: Break Marking GUI output



6.2.1 Pitch Marking

The steps for Semi-Automatic Pitch Marking are mentioned below:

- Detection of Voiced Speech Regions
- Segmentation of Voiced Speech into small segments.
- Pitch Contour Marking.
- Manual Error Correction.

1. Detection of Voiced Speech Regions:

Zero Frequency filtering (ZFF) of the speech is one of the best methods to estimate the voiced epochs. ZFF output contains energy around the zero frequency which is mainly impulse due to excitation. The Procedure of detection of voiced speech region is as follows:

- Extraction of epoch location of clean speech using ZFF [1].
- Add 20dB of white noise to the speech signal.
- Extract the epoch location of noisy speech using ZFF.
- If epoch different between the location obtained in the two cases are less than 2ms, then retain it, otherwise discard the epoch.

- Calculate the instantaneous pitch period using the epoch locations and eliminate the epoch if its pitch period is more than 15 ms.
- Calculate the instantaneous jitter and eliminate the epoch, if jitter is more than 1 ms.

2. Segmentation of Voiced Speech into Small Segments:
Autocorrelation method is used to compute average pitch using a 20 ms frame size with 5 ms frame shift. The discontinuity is calculated by taking difference of the pitch contour.

3. Pitch Contour Marking: Once the segments are obtained, they are marked with one of the following marking: HL (high to low), VHL (very high to low), HVL (high to very low), LH (low to high), VLH (very low to high), LVH (low to very high), and FR (F means flat and R is the average value of the pitch in that segment). One simple way to decide whether a contour is rising or falling can be to see the difference between average value of first and last few samples. But sometimes, spurious detection of pitch values especially at the edges, may lead to incorrect decision. That is why the pitch values are fitted with a line. Fitting is done using linear regression which estimates the coefficients (b_1, b_2) of a polynomial of degree one that fits the values best in a least-squares sense. The estimated line is given in below Equation 6.1.

$$\hat{y} = b_1 x + b_2 \tag{6.1}$$

where, x is a vector = (1, 2, ... N) and N is the length of the segment.
Now the marking is obtained based on the measure of the height value. The height $\hat{y}$ is calculated by taking the difference between the first and the last sample in the fitted line.

$$\hat{y} = y^1 - y^N \tag{6.2}$$

Now if the height value is more than 20, then it is decided that the pitch contour is rising and if it is less than -20, then it is a falling pitch contour. If height is more than 100 or less than -100, then the pitch contour can have a very high or very low pitch value at the edge. Which edge will have the high or low value that is determined by finding the closeness of the pitch values at the edges to the average pitch value (APV) of the speech utterance. If height is in between -20 and 20, the contour is decided to be flat. For a flat pitch contour the average pitch value is calculated and written along with the

flat marking. Based on these conditions, the pitch contours are marked as one of the following: HL, VHL, HVL, LH, VLH, LVH, and FR.

4. Manual Error Correction:

Automatically segmented and marked speech segments are then manually corrected by looking at the output of the automatic marking in the wavesurfer. An error can be in the segmentation boundary or in the pitch contour marking. An error in the segmentation boundary can be of the following types:

- There should not be any segment boundary, but one is detected. In this case an error correction will be to delete the segment boundary.
- There should be a segment boundary, but no segment boundary is detected. In this case an error correction will be to insert a segment boundary.
- The segment boundary is not detected at appropriate position. In this case an error correction will be to shift the position of the segment boundary.

6.2.2 Break Marking:

In order to find the typical length of the word breaks in the speech data, the sentences are first manually analyzed in wavesurfer. Then, the limits of the four types of break indexes B0, B1, B2 and B3 are set. We consider the break index as B0 when the break length (silence portion) between the words is less than 0.080s, B1 when the break length is greater than 0.080s and less than and equal to 0.280s, if the break length is greater than and less than and equal to 0.400s, marked as B2. If the length (duration) of the break is greater than 0.400s, we marked as B3. Then, the automatic break marking is done by detecting the break length (silence portion) of the speech data. The last step is to check the consistency of break marking and do the manual correction with the help of Wavesurfer (version 8.5.2.8).

6.2.3 Semi-Automatic Syllable Labeling

Syllable labeling is the process of partitioning a word into syllables along with time durations. A syllable is a sub-division of a word, typically consisting of a vowel, called the nucleus and the consonant preceding and following the vowel, called the onset and the coda respectively [2]. Most

linguists consider the syllable as an important unit of prosody because many phonological rules and constraints apply within syllables or at syllable boundaries [3]. Apart from the linguistic significance, syllables play an important role in speech synthesis and recognition [3]. One of the major reasons for considering syllable as a basic unit for Automatic Speech Recognition (ASR) system is its better representation and duration stability compared to the phoneme [4]. The syllable was proposed as a unit for ASR as early as in 1975 [5].

The first step of this section is the Phonetic segmentation and alignment of the speech data which is to be syllabified. Phonetic segmentation and alignment determines the time position of the phones of speech corpus based on manual phonetic transcription. It can be done by using the HTK tool HVite with trained phones and manual transcription. The phonetic segmentation and alignment of a sentence is shown in Table 6.1.

Table 6.1: Automatic Phonetic Segmentation and Alignment

Phone Onset(in secs)	**Phone Offset(in secs)**	**IPA(Phone unit)**
0.0000000	0.0500000	sil (silence)
0.0500000	0.1200000	m
0.1200000	0.1700000	ə
0.1700000	0.2100000	n
0.2100000	0.2500000	i
0.2500000	0.3200000	p
0.3200000	0.3800000	u
0.3800000	0.4200000	ɹ
0.4200000	0.4700000	b
0.4700000	0.5400000	i
0.5400000	0.5800000	d^h
0.5800000	0.6400000	a
0.6400000	0.7000000	n
0.7000000	0.8000000	s
0.8000000	0.8600000	ə
0.8600000	0.9200000	b
0.9200000	0.9900000	a

Table 6.2: Automatic Syllabification with time alignment

Syllable Onset(in secs)	**Syllable Offset(in secs)**	**(Syllable)**
0.0000000	0.0500000	sil (silence)
0.0500000	0.1700000	mə
0.1700000	0.2500000	ni
0.2500000	0.4200000	puɹ
0.4200000	0.4700000	bi
0.5400000	0.7000000	d^han
0.7000000	0.8600000	sə
0.8600000	0.9900000	ba

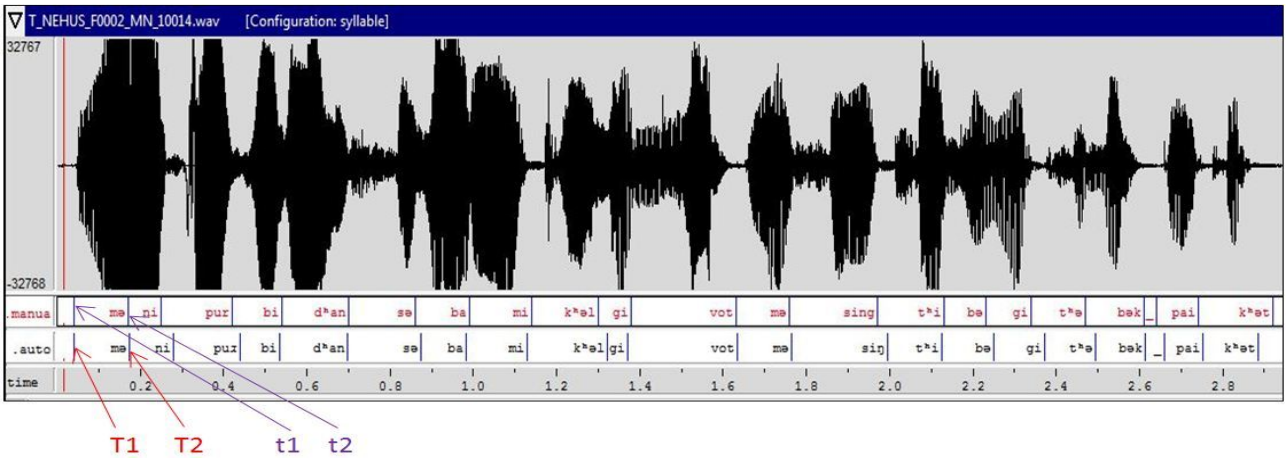


Figure 6.5: Manual and Automatic Syllabification of a sentence displayed using WaveSurfer

The second step is to do the automatic syllabification of the manual phonetic transcription by applying the syllabification rules of Manipuri language. The rules are as follows:

1. A separate nucleus will be produced by each Vowel or diphthong. For example: t^hə-bək

2. If there are two consonants within the two vowels, then the first consonant will be considered as a coda of previous syllabus and the second one as an onset of the next syllable. For example: taŋ-kək.

3. If a single consonant is present in the left side of the nucleus, it will be an onset of the right side syllable. For example: wa-p^həm.

4. If there are three or more consonants between two consecutive vowels, the first consonant will be the coda of the previous syllable while the remaining consonants will be onset of the next syllable. For example : t^hok-k^hɹe.

5. When a consonant is with phone ɹ or y, then that consonant with ɹ or y will be the coda of the syllable. For example : p^ha-k^hɹe.

The third step is the syllable labeling, as shown in Table 6.2. In this step, we extract the time alignment for each syllable using the time alignment of their corresponding Phonetic Segmentation which are done in the first step.

Result Analysis:

In this section, we analyzed the Detection Rate by considering various Time Deviation (W) of the syllable, as shown in the TABLE 6.3. The W of each syllable is calculated by using formula below [6].

Table 6.3: Result Analysis of Semi-Automatic Syllable Labeling System

W per syllable	$\leq$ 20 ms	$\leq$ 30 ms	$\leq$ 40 ms	$>$ 40 ms
Number of syllables Detected	1205	1656	1999	431
Detection Rate	49.5 %	68.1 %	82.26 %	17.7%

$$W = \left| \left(T2 - T1 \right) - \left(t2 - t1 \right) \right| \times 100 \ ms \tag{6.3}$$

Where $T2$ and $T1$ are the syllable offset and onset of automatic syllabification, $t2$ and $t1$ denote the syllable offset and onset of manual syllabification as shown in Fig. 6.5.

50 Manipuri sentences are used for "Result Analysis" which is done by comparing the syllable onset and offset of the manual and automatic syllabification. The total sentences of testing data consists of 2430 syllables. The average Deviation is calculated by using the Equation 6.4:

$$W_{avg} = \sum_{i=1}^{2430} \frac{W_i}{2430} = 25 \ ms \tag{6.4}$$

The Detection Rate is determined by the number of phones detected within the value of a considered W (for example 20ms, 30ms or 40 ms) per total number of phones. The Detection Rate of various Time Deviation "W" is shown in TABLE 6.3.

6.3 Development of Manipuri Phonetic Engine

The Phonetic Engine (PE) is introduced in the literature as a system that captures the phonetic information of the speech signal and transforms it to symbolic form [7,8]. This system produces a sequence of symbols without using any formal knowledge of the language like lexicon, syntactic and semantic. This system obeys the principle of HMM [9,10]. The PE can be used in various applications like keyword detection [11,12], language recognition [13], speaker identification [14], music identification and translation [15,16]. The speech recognition based on phones is very attractive since it is inherently free from vocabulary limitations. Large Vocabulary Automatic Speech Recognition

(LVASR) system's performance depends on the quality of the phone recognizer [17]. The searching speed of database increases, if the phones are used as sub-word units [18].

The processing steps involved in the development of of Manipuri Phonetic Engine are as follows: (A) Task Definition (B) Acoustic analysis (C) Training phase and (D) Training phase.

6.3.1 Task Definition

In the development of the Manipuri Phonetic Engine, the IPA (revision 2005) symbols are used as the sub-word units. The symbols of the IPA have been used for representing the distinct sounds of speech signal [19]. These symbols represent what is spoken rather than what is intended to be spoken. Since IPA symbol captures all distinctive acoustic phonetic characteristics of the speech signal, they can be termed as Acoustic Phonetic Segment (APS) [20]. We merged some of the phonetic units of transcribed data which produce similar sounds, for example, o and ɔ, d͡z and z, etc. After merging the phonetic units of similar sounds, a total of 30 phonetic units including a silence unit are adopted for the development of Manipuri Phonetic Engine. These 29 phonetic units are then assigned 29 ASCII codes, while the silence symbol is denoted by "sil" [21]. Now, by using these 29 ASCII codes along with "sil", we created the basic architecture of our recognizer, which consists of the language model (the task grammar) and the pronunciation model (task dictionary). The HTK recogniser requires the task grammar in Standard Lattice Format (SLF) [22]. Therefore, Task grammar is converted to Task Network which is in SLF.

6.3.2 Acoustic Analysis

The system cannot directly process the speech waveforms [23]. The original waveform has to be converted into a series of acoustic vectors. MFCC feature extraction technique has been used for this purpose. The signal is sampled with a sampling frequency of 16 kHz and segmented into successive overlapping frames of 25 ms each with a frame shift of 10 ms. Each frame is multiplied by a window function (Hamming window). A vector of acoustic coefficients that gives a compact representation of the spectral properties of the frame is extracted from each windowed frame. A 39 coefficient feature vector is then extracted from each frame. Here, each feature vector consists of a log energy, 12 MFCCs,13 delta coefficients and 13 acceleration coefficients respectively. These 39 coefficients are then used to extract vocal tract information of the speaker.

Table 6.4: List of Phonetic units used in Manipuri Phonetic Engine and the Reduced set after merging similar units

Sl No.	Phonetic units	Reduced Phonetic units	Name in ASCII
1	i, ɨ	i	i
2	a	a	aa
3	ə	ae	ae
4	o, ɔ	o	o
5	e	e	ee
6	u, ʉ	u	u
7	n	n	n
8	m	m	m
9	ŋ	ŋ	ng
10	p	p	p
11	b	b	b
12	t	t	t
13	d	d	d
14	k	k	k
15	g	g	g
16	p^h, f	p^h	ph
17	b^h, v	b^h	bh
18	t^h	t^h	th
19	d^h	d^h	dh
20	k^h	k^h	kh
21	g^h	g^h	gh
22	z, $\widehat{\text{dz}}$	z	j
23	s, ʃ	s	s
24	h	h	h
25	w, ʊ	w	w
26	ɹ, ɾ	ɹ	r
27	y	y	y
28	l	l	l
29	$\widehat{\text{ts}}$	$\widehat{\text{ts}}$	ts

6.3.3 Training Phase

For each of the phonemes including silence, a HMM is designed. Each model consists of 5 states. The first and the last states are non-emitting states and the remaining 3 states are active state. The pre-defined prototype along with acoustic vectors and transcription of training data are used by HTK tool HCompV for initialization. This tool is to calculate the global speech mean and variance of HMMs per state.

In the next phase of the development process, the flat start monophones calculated are re-estimated, that is the optimal values for the HMM parameters (transition probability, mean and variance vectors for each observation function) are re-estimated. In our system implementation, re-estimation iteration

are repeated for six times.

6.4 Testing Phase:

The data to be tested are first transformed into a series of acoustic vectors (MFCCs) in the same way as being done during acoustic analysis in the training phase. The acoustic vectors with HMMs definition, task network, dictionary and HMM lists are processed in order to produce the transcription of the test data.

6.5 Experimental Result

Experiments of the phonetic engine were performed on the speech data collected from female and male speakers for Read, Conevrsational and Lecture mode of speech data. During this experiment, three types of speech databases have been created containing for female speakers, male speakers and both male and female speakers for Read and Lecture mode of speech while one type of database for Conversational mode of speech contains data from male speakers only. In read mode of speech, an accuracy of 70.49 % is achieved when the system is trained and tested using database of both male and female speakers together. Further, the system gives accuracies of 74.21% and 72.04% on the separate databases of female and male speakers respectively. For lecture mode of speech, we got an accuracy of 68.62% when the engine is trained and tested with database of both male and female speakers together. An accuracies of 69.71% and 72.5% are achieved with databases of female and male speakers separately. In the Conversation mode, an accuracy of 64.71 % is achieved when the system is trained and tested with the database of speech collected from male speakers. The details are shown in TABLE 6.5.

Table 6.5: Result Analysis of Manipuri Phonetic Engines

Sl. No.	Mode	Types of Speakers	Accuracy in %
1	Conversation	Male only	64.71
2	Lecture	Female only	69.71
3	Lecture	Male only	72.50
4	Lecture	Both male and female	68.62
5	Read	Female only	74.21
6	Read	Male only	72.07
7	Read	Both male and female	70.49

6.6 Summary & Future work

We have successfully developed the three systems namely; Phonetic Engine for Manipuri language is using HTK v3.4., Manipuri Keywords Spotting System using HMM, An HMM based Semi-Automatic syllable labeling System for Manipuri language and also designed GUI for Prosodically guided Manipuri Phonetic Engine. As a future work, we may further analyze by collecting more data of speech from native and non-native speakers of manipuri language from peoples of various communities living in rural and urban areas of Manipur.

6.7 References

1. K. Sri Rama Murty, B. Yegnanarayana, *Epoch Extraction from Speech Signals*, IEEE Transactions on Audio Speech and language Processing, Vol. 16 no. 8. Nov, 2008.

2. Susan Bartlett, Grzegorz kondrak and Colin Cherry, "*Automatic Syllabification with Structured SVMs for Letter-To-Phoneme Conversion*", Proceedings of Association for Computational Linguistics-08: HLT, pages 568-576, Columbus, Ohio USA, June 2008.

3. Susan Bartlett, Grzegorz kondrak and Colin Cherry, "*On the Syllabification of Phonemes*", Human Language Technologies: The 2009 Annual Conference of the North America Chapter of the ACL, pages 308-316, Boulder, Colorado, June 2009.

4. T. Nagarajan, Hema A.Murthy and Rajesh M. Hegde, " *Segmentation of speech into syllable-like units* ", Eurospeech, 2003.

5. Osamu Fujimura, "*Syllable as a unit of speech recognition*", IEEE Trans. Acoust, Speech, Signal Processing, Vol 23, no.1, pp. 82-87, February 1975.

6. ZHang Jin-xi, YU Hong-zhi, MA Ning, LI Zha-yao, "*The Phoneme Automatic Segmentation Algorithms Study of Tibetan Lhasa Words Continuous Speech Stream*", Proceedings of the 2nd International Conference On System Engineering and Modeling, 2013.

7. P. Eswar, *A Ruled-based Approach for Spotting Characters from Continuous Speech in Indian Languages*, Ph.D. thesis, Department of Computer Science and Engineering, IIT Madras, July 1990.

8. S. V. Gangashetty, *Neural Network Models For Recognition of Consonant-Vowel units of speech in multiple languages*, Ph.D. thesis, Department of Computer Science and Engineering, IIT Madras, October, 2004.

9. L.R. Rabiner, *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proc.of the IEEE Vol. 77, Issue 2,pp. 257286,1989.

10. R. Rabiner, and B. H. Huang, *An introduction to hidden markov models*, IEEE Acoust. Speech Signal Processing Mag., pp. 4-16, 1986.

11. P. Schwarz, *Phone recognition based on long temporal context*, Ph.D. thesis, Faculty of Information Technology, Bruno University of Technology, 2008.

12. Rahul. L, Nandakishor. S, Singh L.J and Dutta S.K, *Design of Manipuri Keywords Spotting System using HMM*, NCVPRIG, PP. 1-3, ISBN 978-1-4799-1586-6, 18-21 Dec. 2013.

13. P. Matejka, *Phonotactic and Acoustic Language Recognition*, Ph.D. thesis, Faculty of Electrical Engineering and Communication, Bruno University of Technology, 2009.

14. S. Furui, *50 Years of Progress in Speech and Speaker Recognition Research*, Proceedings of ECTI Transactions on Computer and Information Technology, vol. 1, no. 2, 2005.

15. H. Fujihara, and M. Goto, *Three Techniques for Improving Automatic Synchronization between Music and Lyrics: Fricative Detection, Filler Model, and Novel Feature Vectors for Vocal Activity Detection*, Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 69-72, USA, April 2008.

16. M. Gruhne, K. Schmidt, and C. Dittmar, *Phone recognition in popular music*, Proceedings of 8^{th} International Conference on Music Information Retrieval, Austria, September 2007.

17. Carla Lopes and Fernando Perdigao, *Phone Recognition on the TIMIT Database*, Speech Technologies, pp. 285-356, ISBN 978-953-307-996-7, 2011.

18. Peter S. Cardillo, Mark Clements and Michael S. Miller, *Phonetic Searching vs. LVCSR: How to find what you really want in audio Archives*, International Journal of Speech Technology, no. 5, pp. 922, 2002.

19. International Phonetic Association, *Hand Book of the International Phonetic Association*, Cambridge University Press, the Edinberg Building, Cambridge CB2 2RU, UK 1999.

20. Peri Bhaskararao, *Salient phonetic features of Indian languages in speech technology*, Sadhana, vol. 36, part. 5, pp. 587599, 2011.

21. Salam Nandakishor, Laishram Rahul, S.K Dutta and L. Joyprakash Singh, *Development of Manipuri Phonetic Engine*, Zonal Seminar, The Institute of Electronics and Telecommunication Engineers [IETE], May 3-4, 2013.

22. Steve Young et al., *The HTK Book (for HTK Version 3.4)*, Cambridge University Engineering De- partment, Cambridge, 2009.

23. Mohit Dua, R.K.Aggarwal, Virender Kadyan and Shelza Dua, " *Punjabi Automatic Speech Recognition Using HTK*", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, ISSN 1694-0814, July 2012.

7

Rajiv Gandhi Institute of Technology (RIT) Kottayam

PROGRESS SUMMARY REPORT OF RIT Kottayam

A. General

A.1	Name of the Project	:	**Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages (Malayalam & Kannada)**
	Our Reference Letter No	:	11(6)/2011-HCC(TDIL) dated 23-12-2011
A.2	Executing Agency	:	RIT Kottayam
A.3	Chief Investigator with Designation	:	Dr. Leena Mary Professor, Dept. of ECE, RIT Kottayam
	Co-Chief Investigators with Designation	:	Mr. Riyas K. S., Mr. Anish Babu K. K Asst. Professor, Dept. of ECE, RIT Kottayam
A.4	Project staffs with Qualification	:	Jubin James Thennattil - M.Tech Anil P. Antony - B.Tech
A.5	Total Cost of the Project as approved by DIT	:	50.6 Lakhs (Total)
	i) Original	:	27.6 Lakhs (for the first year)
		:	23.0 Lakhs (for the second year)
	ii) Revised, if any	:	
A.6	Date of starting (Indicate date of first sanction)	:	23/12/2011
A.7	Date of Completion		
	i) Original	:	23/12/2013
	ii) Revised, if any	:	31/03/2015
A.8	Date on which last progress report was Submitted	:	28/02/2015

B. Technical

B.1 Works Progress.(given details in technical report) :

- Database collection in three different modes:
  - i. Read speech : 15 hrs (Malayalam), 10 hrs (Kannada)
  - ii. Lecture mode : 07:30 hrs (Malayalam), 10 hrs (Kannada)
  - iii. Conversational speech : 07:30 hrs (Malayalam), 10 hrs (Kannada)
- Transcription using IPA chart: 10 hrs (Malayalam), 10 hrs (Kannada)
- Acquisition of prosody
- Development of Phonetic Engine
- Method for audio search using phonetic and prosodic labels

C. Financial

Consolidated statement of expenditure (in Rupees)

Sl. No	**Sanctioned Heads**	**Funds Received**	**Expenditure incurred**	**Balance**
		A	**B**	**A−B**
1.	Capital Equipment	6,00,000.00	5,76,091.00	23,909.00
2.	Data Collection	10,00,000.00	10,57,805.00	-57,805.00
3.	Consumable stores	3,00,000.00	2,10,284.50	89,715.50
4.	Manpower	16,00,000.00	18,74,952.00	-2,74,952.00
5.	Travel	2,00,000.00	2,13,491.00	-13,491.00
6.	Workshop & Training	2,00,000.00	1,65,266.00	34,734.00
7.	Contingency	3,00,000.00	3,44,875.50	-44,875.50
8.	Coordination & Management	2,00,000.00	1,23,808.00	76,192.00
9.	System Integration	0.00	0.00	0.00
10.	Sub Total	44,00,000.00	45,66,573.00	-1,66,573.00
11.	Over Head (15%)	6,60,000.00	4,93,427.00	1,66,573.00
12.	**Total Budget**	50,60,000.00	50,60,000.00	0

D. Project Outcomes

Papers Published:

- Leena Mary (2012). Extraction and representation of prosodic features for speaker, speech and language recognition, *Springer briefs in Electrical and Computer Engineering*, ISBN 978-1-4614-1158-1, 2012, Publisher: Springer Netherlands.

- Anish Augustine, Riyas K. S, Leena Mary, "Phonetic transcription and labelling of prosody for automatic speech recognition", *Proc. National Technological Congress (NATCON13)*, Feb 2013.

- Sreejith A, Leena Mary, Riyas K.S. , Aju Joseph, Anish Augustine, "Automatic prosodic labeling and broad class Phonetic Engine for Malayalam", *Proc. IEEE International Conference on Control Communication and Computing (ICCC)*, Dec 2013.

- Gayathri M. R, Anil P. Antony, and Leena Mary, "Automatic syllabification of speech signals", *Proc.* 15^{th} *National Conference on Technological Trends (NCTT)*, Thiruvananthapuram, Aug. 2014.

- Deekshitha G, and Leena Mary,"Broad phoneme classification using signal based features" in *Proc. International Journal on Soft Computing (IJSC)*, Vol. 5, No. 3, November 2014.

- Deekshitha G, Jubin James Thennattil and Leena Mary,"Implementation of Automatic Segmentation of Speech Signal for Phonetic Engine in Malayalam" in *Proc. International Journal of Engineering and Technical Research (IJERT)*, Vol. 2, Issue.11, November 2014.

- Shridhara, M.V., Banahatti, B.K., Narthan, L, Karjigi, V., Kumaraswamy. R, Development of Kannada speech corpus for prosodically guided phonetic search engine, *Proc. IEEE Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pp 1-6, Nov. 25-27,2013, Gurgaon, India.

- Deekshitha G, Jubin James Thennattil, and Leena Mary, "Segmentation of continuous speech for broad phonetic engine", in *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (IEEE ICECCT 2015)*,SVS College of Engineering, Coimbatore, pp 1147 - 1151, March 2015.

- Deekshitha G., and Leena Mary, "Prosodically Guided Phonetic Engine" , in *IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (IEEE SPICES)*, February 2015, NIT Calicut.
- Jubin James Thennattil, and Leena Mary, "Real time phonetic engine for large vocabulary continuous speech in Malayalam", in *IETE Journal of Research* (under review).
- Leena Mary, Anil P. Antony and S. R. M. Prasanna, "Automatic syllabification of speech signals using short time energy and vowel onset points", in *Circuits, Systems and Signal Processing, Springer* (under review)

Development of Database

- Read speech: 15 hrs (Malayalam), 10 hrs (Kannada)
- Lecture mode: 07:30 hrs (Malayalam), 10 hrs (Kannada)
- Conversation: 07:30 hrs (Malayalam), 10 hrs (Kannada)

Transcription of Database

- Read speech: 5 hrs (Malayalam), 5 hrs (Kannada)
- Lecture mode: 2:30 hrs (Malayalam), 2:30 hrs (Kannada)
- Conversation: 2:30 hrs (Malayalam), 2:30 hrs (Kannada)

Tools & Systems Developed

- Phonetic engine for Malayalam language
- Phonetic engine for Kannada language
- Acquisition of Prosody - Automatic Syllabification of Speech Signal
- Acquisition of Prosody - Automatic Breakmarking of Speech Signal
- A coarse audio search method using phonetic and prosody labels

Appendix 7.1

Detailed Technical Report of RIT Kottayam

Detailed Technical Report of RIT Kottayam

Development and availability of spoken language corpora in regional languages is of utmost importance for a multicultural and multilingual country like India. The issues of regional bias, accent, unique style and diversity associated with each geographical region and language will have a significant effect on the performance of speech recognition/synthesis systems. Collection of speech data in Malayalam and Kannada language for prosodically guided phonetic search engine and the issues involved in transcription are explained in this report. The speech corpus consists of data in three different contexts namely, read mode, conversation mode and extempore mode. A four layered transcription namely, phonetic transcription using IPA symbols, syllabification, pitch marking and break marking is done. A baseline recognition system for Malayalam and Kannada language is built using HTK for the data collected in different modes and the results are presented.

The tasks addressed at Rajiv Gandhi Institute of Technology in connection with the project are the following:

9.1 Database collection & transcription

9.2 Identification and marking of prosodic events

9.3 Automatic prosody marking

9.4 Broad phonetic labelling using signal level features

9.5 Development of phonetic engine

9.6 Coarse method for audio search

9.7 Details of training programme conducted

9.8 Summary and future work

Progresses of each task are described below:

7.1 Database collection & transcription

7.1.1 Data Collection in Malayalam and Kannada

Data has been collected in three different modes namely read speech, extempore speech (lecture mode) and conversational speech. Details of data collected in Malayalam by RIT Kottayam and in

Kannada by SIT Tumkur are tabulated.The conversational data were recorded in the field, from four different regions of Kerala for Malayalam. Recording has been done using Zoom H4N voice recorder with a sampling frequency of 48kHz and represented using 16 bits per sample. The details of overall data collection in each mode from different regions is tabulated below.

Table 7.1: Details of data collected in Malayalam

Mode	**Total Duration**	**No. of Spkrs.**	**No. of Male Spkrs.**	**No. of Female Spkrs.**
Read Speech	15 hrs	33	18	15
Lecture mode	07:30 hrs	20	14	06
Conversation	07:30 hrs	49	33	16

No. of regions data has been collected	Regions
5	Trivandrum, Kottayam, Alappuzha, Thrissur, Kannur

Table 7.2: Details of data collected in Kannada

Mode	**Total Duration**	**No. of Spkrs.**	**No. of Male Spkrs.**	**No. of Female Spkrs.**
Read Speech	10 hrs	26	15	11
Lecture mode	10 hrs	10	08	02
Conversation	10 hrs	26	18	08

No. of regions data has been collected	Regions
5	Bangalore, Mangalore, Mysore, Dharwad, Belgaum

7.1.2 Transcription to IPA symbols

The data obtained is divided into smaller chunks of duration 2-5 minutes in each mode for further processing, such as transcription, syllabification, pitch marking and break marking. Once chunking gets over, the data is ready for being transcribed. While transcribing, the signal is carefully listened and looked into so as to minimize transcription error as much as possible. Transcription has been done using the International Phonetic Alphabet (IPA) chart. Details of IPA transcription is provided in the table.

Table 7.3: Details of IPA transcription in Malayalam

Mode of Speech	Duration transcribed using IPA symbols
Read speech	5 hrs
Extempore (lecture mode)	2 hrs 30 mins
Conversational speech	2 hrs 30 mins

Table 7.4: Details of IPA transcription in Kannada

Mode of Speech	Duration transcribed using IPA symbols
Read speech	5 hrs
Extempore (lecture mode)	2 hrs 30 mins
Conversational speech	2 hrs 30 mins

7.2 Manual prosody marking

Prosody is of interest to automatic speech recognition (ASR), as it is important for human speech recognition. The role of prosody is particularly important in spontaneous speech. Conversational speech contains large amount of prosodic variation, which seems to co-occur with greater acoustic variability. Researchers have long hypothesized that prosody could be useful in improving computer recognition of speech. However, prosody has been used to only a small extent, though successful applications in ASR are growing.

All spoken utterances can be considered as sequence of syllables which constitute a continual rhythmic alternation between opening and closing of mouth while speaking. Syllable of CV type provides an articulatory pattern beginning with a tight restriction and ending with an open vocal tract, resulting some rhythm that is especially suited both to the production and the perception mechanisms. It is demonstrated that the tonal events are aligned to the segmental events such as onset and/or offset of a syllable. Therefore, syllable appears to be a natural choice for the basic unit for representing prosody.

In this work, we have assumed that prosody is manifested at syllabic level in terms of tonal variations and breaks. In order to acquire prosodic knowledge, the following three automatic prosodic transcription tasks are attempted:

(i) Automatic detection of syllable boundaries

(ii) Automatic labelling of pitch accents

(iii) Automatic detection and labelling of break indices

In order to evaluate the three above mentioned automatic prosody transcription, it is necessary to have manually labelled prosodic data. Manual prosodic transcription for all 12 languages is done as per the following steps:

(i) 300 sentences (approximately 1 hour), i.e., 100 sentences in each in read, lecture and conversation mode, is separated for this transcription.

(ii) Time stamps are marked corresponding to syllable boundaries

(iii) Pitch accents corresponding to local variations are marked with VL, L, H and VH.

(iv) Break Indices B1 (inter word), B2 (phrase break) and B3 (sentence break) are marked.

7.2.1 Syllabification

Syllable is a larger unit than a phoneme. A syllable is composed of a central peak of sonority (usually a vowel), and the consonants that cluster around this central peak. Syllable has a nucleus, normally a vowel sound. A word can be divided into syllables with a nucleus and it should follow the phonotactics of the language. In syllabification procedure, syllable boundaries are manually marked and labelled using wave surfer.

7.2.2 Break marking

Syllable boundaries are marked with any of the four labels namely B_0, B_1, B_2, B_3 where B_0 corresponds to syllable boundary marking which do not have a physical break in the waveform. B_3 corresponds to long pauses such as sentence break. Break marking indicate the duration of pause between words. B_0 is the smallest break where adjacent syllables are joined together and no physical break is present.

7.2.3 Pitch marking

Pitch is a perceptual attribute of sound which can be described as a sensation of the relative altitude of sound. The physical correlate of pitch is the fundamental frequency (F_0) determined by the rate of

vibration of the vocal chords. The ensemble of pitch variations in the course of an utterance is defined as intonation. The direction of F_0 change, either rising or falling, is determined by the phonological patterns of the constituent words. Four labels namely Very High (VH), High (H), Low (L) and Very Low (VL) are used for marking pitch variations. Flat segment pitch is marked with its absolute F_0 value.

7.3 Automatic prosody marking

Automatic prosody marking consists of automating the process of break marking, pitch marking and syllabification

7.3.1 Automatic Break marking

To find the break marking such as B_0, B_1, B_2 and B_3, speech/non speech regions are identified first. Each frame of 10 ms is classified as speech or non speech based on average short time energy (SE), spectral flatness measure (SFM) and most dominant frequency (MDF) calculated for that frame. For non speech, average short time energy will be less compared to speech. Spectral flatness measure is the measure of white noise in a spectrum. Most dominant frequency gives highest frequency within the frame, it will be invariably less for non speech frame. Above three features are used to classify each frame as speech or non speech giving three independent decisions. Final speech or non speech decision is taken based on majority voting. Breaks are marked as B_0, B_1, B_2 and B_3 depending on the duration of non speech region around the syllable boundary.

7.3.2 Automatic Pitch Marking

Pitch values at start, end and maximum value positions are compared with the reference average (RefAvg) and labelled as L, H, VH, VL or Favg based on the following strategies:

1. Low (L), if pitch value F_0 is such that (0.5*RefAvg $< F_0 <$ RefAvg)
2. High (H), if (RefAvg $< F_0 <$ 1.5 * RefAvg)
3. Very Low (VL) if F_0 <0.5 RefAvg
4. Very High (VH) if $F_0 >$ 1.5 *RefAvg

5. If start pitch value, end pitch value and the F_{max} are equal or nearly equal within a tolerance range, then it can be labeled as F_{avg} which stands for the average value of pitch for that syllable

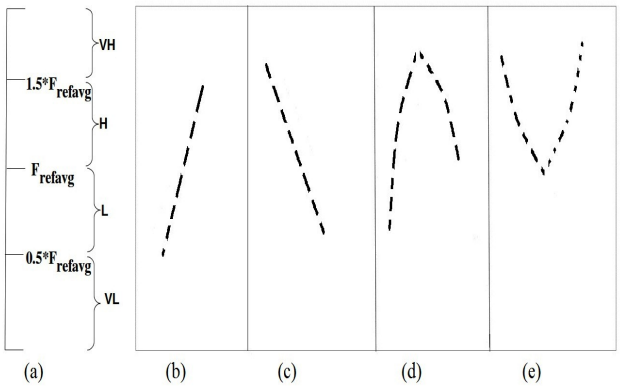


Figure 7.1: Pitch contour categories considered for pitch marking

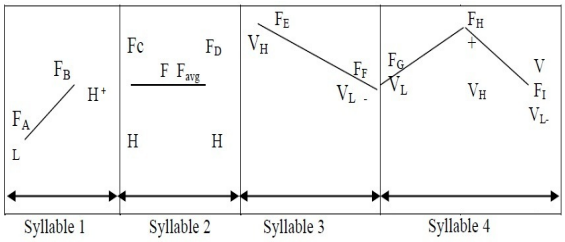


Figure 7.2: Illustration of automatic pitch marking algorithm

We have considered mainly four types of pitch contour 'rising' as in Fig. 9.1(b), 'falling' as in Fig. 9.1(c), 'rising and falling' as in Fig. 9.1(d) and 'falling and rising' as in Fig. 9.1(e). We are classifying pitch contours into above four categories. As a first step, pitch values are median filtered for smoothing the contour. Pitch marking is done for each syllable using the boundary information derived via manual syllabification procedure. A reference average of pitch (F_{refavg}) is computed for every 5 second moving window and pitch levels are marked as VL, L, H and VH based on this average pitch. Maximum (F_{max}) and minimum (F_{min}) value of pitch and its position are computed for each syllable region. If position of F_{min} coincides with starting point of pitch contour and position of F_{max} with ending point of pitch contour then pitch contour is considered as 'rising' contour. If position of F_{max} coincides with starting point of pitch contour and position of F_{min} with ending point of pitch contour, then it is considered as 'falling' contour. If (F_{max}) happens to be somewhere in the middle, it is 'rising and falling' contour. If start pitch value, end pitch value and the F_{max} are equal or nearly equal within a tolerance range, then it can be labelled as F_{avg} which stands for the average value

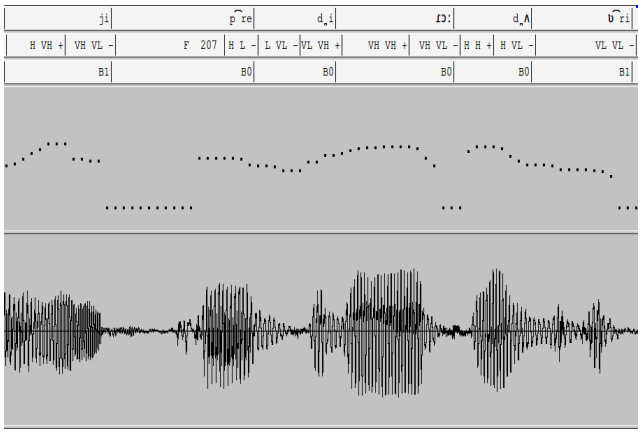


Figure 7.3: Automatic labeling of pitch marking and break marking (shown above pitch contour and speech waveform)

of pitch for that syllable. The pitch values at the end position is obtained by taking the difference between F_{start} and F_{end} points. For falling contour

(i) If $0.2*F_{refavg} > (F_{start} - F_{end}) \geq 0.1*F_{refavg}$, then end point is one level below the starting point.

(ii) If $0.3*F_{refavg} > (F_{start} - F_{end}) \geq 0.2*F_{refavg}$, then end point is two level below the starting point.

(iii) If $(F_{start} - F_{end}) \geq 0.3*F_{refavg}$, then end point is two level above the starting point.

Algorithm can be explained with the help of Fig. 9.2. The point A is the starting point of pitch contour and B is the end point for syllable 1. Syllable 1 belongs to rising category since F_A coincides with F_{min} and F_B coincides with F_{max}. F_A is labeled as L as F_A falls between F_{refavg} and $0.5*F_{refavg}$. The end point F_B is labelled based on on the difference between F_A and F_B. In this figure the difference is assumed to be greater than $0.1*F_{refavg}$ then F_B is labeled as one level above L which is H. The label ends with a positive sign (+) to indicate the rising category. For syllable 2, the start pitch F_C, end pitch F_D and F_{max} are nearly equal therefore it is labelled as F F_{avg}. For syllable 3, F_E is labeled based on difference between F_E and F_D. End point F is labelled based on the difference between F_E and F_F. In the case of syllable 4, F_{max} lies nearly in the middle of starting and end point so it is classified as rising and falling. We sub segment the syllable boundary up to F_{max} and classify from starting point to position of F_{max} as rising contour and label it similar to rising contour. The remaining part of syllable 4 is classified as falling contour and labelled similar to falling contour.

7.3.3 Automatic syllabification

Automatic syllabification is the segmentation of a sentence/word into syllables by detecting the boundaries automatically. Some of the features useful for automatic syllabification are amplitude of a speech signal, short time energy, breaks in F0 contour and locations of vowel onset points. Utilizing some of these features, a methodology is proposed for automatic syllabification, which utilizes short time energy and vowel onset points. The block diagram shown in Figure 7.4 illustrates the proposed methodology for automatic syllabification.

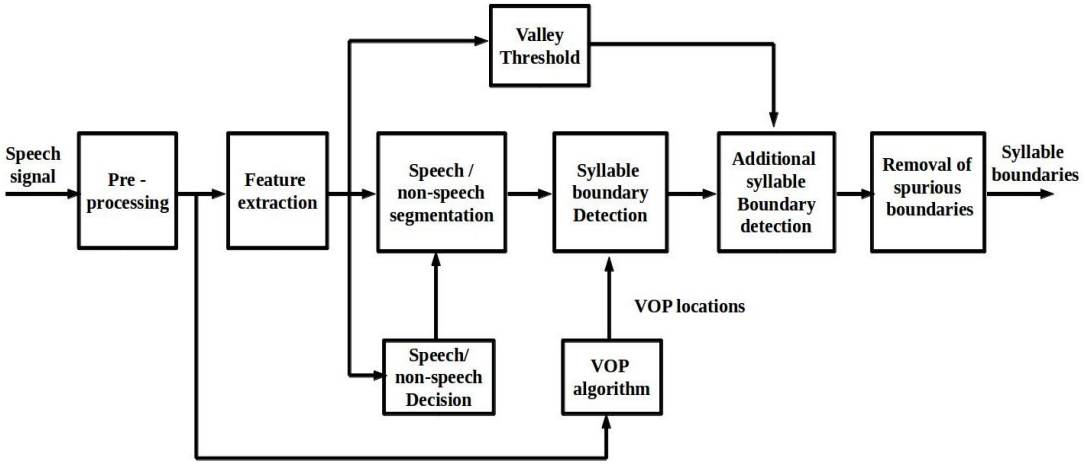


Figure 7.4: Block schematic illustrating the proposed methodology for automatic syllabification

7.3.3.1 Methodology

The proposed automatic syllabification procedure has the following steps:

(i) Preprocessing

Speech signal is normalized to take care of amplitude variations in recording, by which maximum and minimum amplitudes are limited to $\pm$ 1. Sampling rate is changed to 1kHz.

(ii) Feature extraction

Some of the features useful for automatic syllabification are amplitude of a speech signal, short

time energy, pitch breaks. The short time energy (STE) is computed with 20ms frame size and 10 ms frame shift using Hamming window. Most dominant frequency in the spectrum (MDF) is computed by taking the absolute value of maximum of FFT of signal framed to 20ms with 10ms.The pitch extracted gives the voicing information.

(iii) Speech/Non speech segmentation and marking of long silence/pause regions

Speech/nonspeech detection is performed using STE , MDF and voicing information. For each frame, speech/nonspeech decision is taken based on whether the feature has value greater than a threshold and whether the frame is voiced/unvoiced. This results in three different decisions for each frame based on STE, MDF and voicing information. Majority decision among the three is chosen as the final speech/nonspeech decision . Sufficiently long nonspeech regions (approx. above 100 msec) are marked as silence. Nonspeech regions of shorter durations are not marked.

(iv) VOP Detection

VOP detection is performed using the change of strength of excitation represented in Hilbert envelope of linear prediction residual. First the speech signal is processed in blocks of 20 ms with a shift of 10 ms. For each 20 ms block, 10^{th} order LP analysis is performed to estimate the linear prediction coefficients (LPCs) . The time-varying inverse filter is constructed using these LPCs. The speech signal is passed through the inverse filter to extract the LP residual signal. The time varying nature of excitation source characteristic is further enhanced by computing Hilbert envelope of the LP residual [4]. For every 5 ms block with one sample shift, the maximum value of the Hilbert envelope of LP residual is noted to construct smoothed excitation contour.

The change in the excitation characteristics at the VOP is detected by convolving the smoothed excitation contour with a first order Gaussian differentiator of length 100 ms and standard deviation as one sixth of the window length. This convolved output is termed as VOP evidence plot. In the VOP evidence plot relative maxima occur at the instants where the amplitude in the Hilbert envelope of the LP residual starts rising sharply.The peaks in the VOP evidence plot represent the locations of the VOPs and are automatically located by finding the maximum value between two successive positive to negative zero crossing with some threshold to eliminate

the spurious ones [5]. The VOP evidences are shown in Figure 7.5 (c).

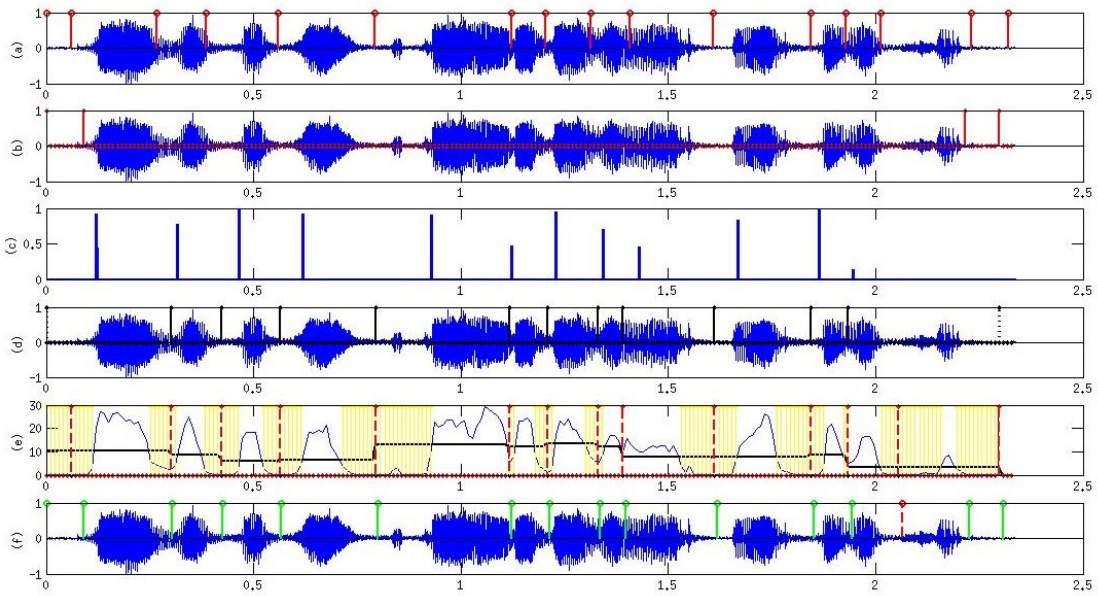


Figure 7.5: Automatic syllable boundary detection for a Malayalam sentence in read mode (a) Speech waveform with manually marked syllable boundaries (b) Speech/non speech segmentation (c) Locations of VOP (d) Boundaries detected using VOP (e) Short time energy (horizontal line shows valley threshold, vertical lines showing valleys detected & dashed vertical lines denote additional syllable boundaries detected using energy valleys) (f) Final set of syllable boundaries after spurious removal.

(v) Detection of syllable boundaries using VOP

If there are more than one VOP between two consecutive speech region, then it an indication of syllable boundaries within this region. Then search is performed for minimum STE position in energy contour between middle point of two consecutive VOPs to the second VOP. This minimum STE positions are marked as syllable boundaries as shown in Figure 7.5 (d).

(vi) Additional syllable boundary detection using energy valleys

For identifying the additional syllable boundaries, a higher threshold is set for STE. For each region of STE below this threshold, valley point (minimum STE position) is identified as syllable boundaries as shown in Figure 7.5 (e) in dashed lines [6]. If there is more than one point with minimum STE in a valley region, then one in the middle position is selected.

(vii) Removal of spurious boundaries

Spurious boundaries within 60ms range are removed first. For this we take speech non speech segments as reference and boundaries detected with valleys as input.Then removing boundaries with distance less than 60ms but keeping the boundary with lowest energy. Then combing the resultant boundaries obtained from the first spurious removal with the syllable boundaries detected using VOP. After this a second level spurious removal is carried out to remove very close spurious boundaries within 40 ms range.

7.3.3.2 Performance Evaluation

Performance of the proposed automatic syllabification method is evaluated by comparing the automatically detected syllable boundaries with the manually marked ones. Sentences from each mode in five languages are used for evaluating the above algorithm. The time stamps of syllable boundaries are manually marked for these sentences. The metrics used for performance evaluation are as follows :

(i) Percentage of detection accuracy : It gives the number of syllable boundaries detected with deviations within ±40 ms and ±50 ms, with respect to manual boundaries.

(ii) Percentage of missed syllable boundaries : It refers to the boundaries present in manual syllabification, but not present in automatic syllabification within ±40ms and ±50ms.

(iii) Percentage of spurious boundaries : It refers to the boundaries present in automatic syllabification, but not present in manual syllabification within ±40ms and ±50 ms.

Table 7.5: Performance evaluation of automatic syllabification for Read mode

Language	Detection Accuracy (in %)		Missed Syllables (in %)		Spurious Syllables (in %)	
	±40 ms	±50 ms	±40 ms	±50 ms	±40 ms	±50 ms
Bengali	74.93	82.05	25.07	17.95	14.09	10.28
Gugarati	75.17	82.22	24.83	17.78	23.56	16.28
Malayalam	80.81	84.86	19.19	15.14	18.68	14.55
Marati	71.51	76.29	28.49	23.71	30.58	25.71
Odiya	70.04	76.22	29. 96	23.78	25.69	19.02

Table 7.5, 7.6 and 7.7 summarises the results of performance evaluation for read, extempore and conversation mode respectively. Performance in terms of detection accuracy in % (within ±40 ms & ±50 ms), missed syllables in % and spurious syllables in % for five languages in read mode is given

Table 7.6: Performance evaluation of automatic syllabification for Extempore mode

Language	Detection Accuracy (in %)		Missed Syllables (in %)		Spurious Syllables (in %)	
	±40 ms	±50 ms	±40 ms	±50 ms	±40 ms	±50 ms
Bengali	78.49	86.35	21.50	13.65	18.25	10.14
Gugarati	65.38	75.58	34.92	24.42	37.42	27.36
Malayalam	76.69	80.73	23.31	19.27	14.90	10.28
Marati	42.86	52.20	57.14	47.80	63.08	54.97
Odiya	72.67	75.34	27.33	24.66	26.99	19.87

Table 7.7: Performance evaluation of automatic syllabification for Conversation mode

Language	Detection Accuracy (in %)		Missed Syllables (in %)		Spurious Syllables (in %)	
	±40 ms	±50 ms	±40 ms	±50 ms	±40 ms	±50 ms
Bengali	69.37	77.46	30.62	22.53	27.35	18.80
Gugarati	67.02	74.99	32.98	25.01	40.02	32.95
Malayalam	71.67	74.86	28.33	25.14	14.72	10.63
Marati	53.92	65.08	46.08	34.92	55.76	46.81
Odiya	65.34	70.08	34.66	29.99	21.67	16.50

in Table 7.5. Performance in terms of detection accuracy in % (within ±40 ms & ±50 ms), missed syllables in % and spurious syllables in % for five languages in extempore mode is given in Table 7.6. Performance in terms of detection accuracy in % (within ±40 ms & ±50 ms), missed syllables in % and spurious syllables in % for five languages in conversation mode is given in Table 7.7.

7.4 Broad phonetic labelling using signal level features

In this work, broad phoneme identification is attempted using signal level feature. For broad phoneme identification, the features such as voicing, nasality, laterality, frication, trilling, vowel formants etc. are studied. From this study, we have come up with a set of feature vectors capable of performing broad phoneme classification. We have broadly divided into six classes and they are Vowels(V) , Nasals(N), Stops(S), Fricatives(F), Approximants(A) and Silence(S).

The useful features identified are Voiced/Unvoiced decision, Zero Crossing Rate (ZCR), Most Dominant Frequency (MDF), Spectral Flatness Measure (SFM), Short time energy, First three formants and Magnitude at the dominant frequency. Signal is windowed for a 20ms duration with 10ms overlap. For each frame, features are computed. These features are normalized and is applied to a neural network classifier as input and broad phoneme class label as output.

Table 7.8: Broad phoneme classes

Sl. No	Type	Symbol	IPA Symbols
1	Vowels	V	a i u e o aː iː uː eː oː
2	Nasals	N	m n̪ n ɳ ɲŋ
3	Stops	P	b d̪ ɖ dʒ g b^h d̪^{h} ɖ^h dʒ^h g^h p t̪ ʈ tʃ k p^h t̪^{h} ʈ^h tʃ^h k^h t
4	Fricatives	F	f s ʂ ʃ h
5	Approximants	A	ʋ ɻ j l ɭ ɾ r
6	Silence	S	

7.5 Development of phonetic engine

The continuous phoneme recognition system (phonetic engine) is developed using HTK (Hidden markov model Tool Kit). The Hidden Markov Models (HMMs) are developed for read, lecture and conversation modes for Malayalam and Kannada.

First step was data preparation. This phase consists of labelling of speech signal. The labelled file saved in .lab format is a simple text file with time stamps of speech data. The speech .wav files were chopped data files into smaller chunks of about 1 to 8 second and also to obtain their respective transcription. The speech wave files and their transcription with time stamps are considered. To build HMM using HTK the IPA symbols present in transcription has to be converted into ASCII symbols. For example: **aːkaːʃəʋaːni** in **IPA** notation is converted into **akashavani** in **ASCII** notation.

The speech recognition tools cannot process the speech waveforms directly, hence it has to be represented in more compact and efficient way. Hence speech waveform is converted into a series of vectors. MFCCs are used for feature extraction. Configuration file (.conf) is a text file which specifies various configuration parameters such as format of speech file, sampling frequency, frame size, frame shift, number of Mel Frequency Cepstral Co-efficients (MFCCs) etc. The speech signal is pre-emphasized by pre-emphesis factor of 0.97. The speech signal is hamming windowed with a frame size of 25 ms duration and 10 ms overlap. For each frame MFCCs are computed. Here 13 MFCCs +13 Δ +13Δ Δ are considered. Hence each feature vector with dimensionality 39 is considered.

Each phone is modelled using 5 state HMM model. The HMM is trained with 5 states, 32 Gaussian mixtures. The HMM means are initialized to zero, variance to 1 and trained with many iterations.

Almost 75% of data is used for training of HMM phoneme models and rest 25% is used for testing purpose.

7.5.1 Development of phonetic engine for Malayalam language

Analysis of transcribed data has been done by tabulating the various IPA symbols used. It was found that a large number of symbols are much less frequent. The number of occurrences of them is very small such that there is lack of examples for model training. So such symbols are further mapped to more frequently occurring similar symbols based on perceptual similarity. Considering frequency of occurrence, we have finalized 40 classes of phonemes including silence. The silence regions in speech signals were transcribed using a special character '-'. Here, the normalized frequency is obtained by dividing the count of phoneme by total count of phonemes. Aspirated consonants, which are very less frequent are mapped to their unaspirated counterparts. Long vowels, and some of the fricatives are not merged since they have much distinct perceptual properties.

Initially, out of four hours of transcribed read mode data available, training was done for 75% of transcribed data and remaining part is used for testing. Training and testing data was selected to balance the gender characteristics. The phonetic engine was made using HTK. The performance of the phonetic engine was evaluated. It was observed that overall phone recognition correctness was 44.24% and overall phone recognition accuracy was reduced to 26.34%. The first phoneme, /a/ has confusion with its long version as well as with other more frequent vowels. Second row show that long vowel /a:/ has most confusion with /a/. Observing rows till tenth, we can infer that long vowels like /ee/, /oo/, /E/ shows much confusion with their short vowel counterparts, due to lack of sufficient data. Phonemes which are stops and fricatives like /b/, /f/, /t/ have less accuracy due to the lack of sufficient data for training. Vowels and silence have comparatively good accuracy. Considering rows 13, 16, 19, 22, 23 and 27, we can see that nasals like /nj/, /ng/ etc. has confusion with other nasals. Similarly, in rows 32, 33 and 34, we can observe that fricatives like /s/, /S/, /sh/ shows similar confusion, due to limited training data.

Table 7.9: Details of mapping of IPA symbols

No	Phonetic Symbols in IPA	Mapped IPA symbol	ASCII symbol	Normalised frequency
1	a ʌ ɐ æ ɑ	a	a	0.1346
2	aː ʌː ɐː æː ɑː	aː	aa	0.0392
3	i y ɪ ʏ ɨ	i	i	0.0858
4	iː yː ɪː ʏː ɨː	iː	ee	0.0042
5	u ʊ ɯ	u	u	0.0445
6	uː ʊː ɯː	uː	oo	0.0041
7	e ø ɘ ə ɛ ɜ	e	e	0.0825
8	eː øː ɘː əː ɛː ɜː	eː	E	0.0106
9	o ʉ ɤ ɵ	o	o	0.0084
10	oː ʉː ɤː ɵː	oː	O	0.0068
11	k k^h	k	k	0.0430
12	g ɢ g^h ʔ	g	g	0.0125
13	ŋ ɴ	ŋ	ng	0.0093
14	ʧ c $ʧ^h$	ʧ	ch	0.0096
15	ʤ $ʤ^h$ ɟ dʒ $dʒ^h$	dʒ	j	0.0055
16	ɲ	ɲ	nj	0.0050
17	ʈ $ʈ^h$	ʈ	T	0.0122
18	ɖ $ɖ^h$	ɖ	D	0.0196
19	ɳ	ɳ	N	0.0135
20	t̪ $t̪^h$	t̪	th	0.0418
21	d̪ $d̪^h$	d̪	d	0.0210
22	n	n	n	0.0289
23	n̪	n̪	$n\_$	0.0365
24	p p^h	p	p	0.0283
25	f ɸ β	f	f	0.0015
26	b b^h	b	b	0.0095
27	m	m	m	0.0466
28	j ʎ	j	y	0.0374
29	ɾ	ɾ	r	0.0250
30	l ʟ	l	l	0.0288
31	v ʋ	ʋ	v	0.0225
32	ʃ ɕ ç	ʃ	S	0.0080
33	ʂ ʒ ʐ ɻ	ʂ	sh	0.0060
34	s z	z	s	0.0235
35	h ɦ ħ χ	h	h	0.0043
36	ɭ ɰ	ɭ	L	0.0154
37	r ʀ	r	rr	0.0324
38	ɻ ɾ	ɻ	zh	0.0023
39	t	t	t	0.0092
40	-	-	-	0.0203

Table 7.10: Performance of phoneme models

No	Phoneme in ASCII	Normalised frequency	% of corectness	confusable phonemes (% of misclassification in bracket)
1	a	0.1346	55.5	aa(7.28), e(12.43), i(3.02)
2	aa	0.0392	68.4	a(13.17)
3	i	0.0858	62.3	e(8.26), y(4.24)
4	ee	0.0042	13.1	i(5,16), r(3.34), e(2.9)
5	u	0.0445	53	e(10.16), a(6.48)
6	oo	0.0041	10.9	u(30.23)
7	e	0.0825	52.9	a(10.23)
8	E	0.0106	27.6	e(18.58), i(12.69)
9	o	0.0084	23	u(15.8), a(14.78)
10	O	0.0068	30.6	a(14,65), o(8.18)
11	k	0.0430	57.2	y(4.2), th(4.01)
12	g	0.0125	22.4	k(9.67)
13	ng	0.0093	21.3	n(13.7), e(9.5)
14	ch	0.0096	45.3	y(9.3), s(5.627), S(5.19)
15	j	0.0055	42.3	y(12.2), r(10.30
16	nj	0.0050	16	n(12.23), y(9.57)
17	T	0.0122	40	th(9.26), k(9.05)
18	D	0.0196	55.9	r(5.37)
19	N	0.0135	30.9	m(7.54), n(9.12)
20	th	0.0418	37.7	k(9.52), p(6.76)
21	d	0.0210	30.9	D(7.02), rr(5.46)
22	n	0.0289	41.2	m(6.3), y(5.5)
23	$n\_$	0.0365	33.3	n(9.07), m(8.5)
24	p	0.0283	47.6	k(8.17), th(7.77)
25	f	0.0015	5.8	s(15.9), a,p(8.69)
26	b	0.0095	34.1	k(7.225), d(6.07)
27	m	0.0466	55.4	$n\_$(5.89), n(4.41)
28	y	0.0374	64.5	i(5.16), r(3.34)
29	r	0.0250	51.7	rr(5.2), y(5.9)
30	l	0.0288	62.8	L(4.86), e(5.34)
31	v	0.0225	34.3	$n\_$(5.28), D(4.05)
32	S	0.0080	41.8	sh(11.10, s(18.13), ch(7.58)
33	sh	0.0060	41.1	S(15.52), s(12.33)
34	s	0.0235	77.3	th(6.1), S(1.9)
35	h	0.0043	37.3	k(8.86), y(8.23)
36	L	0.0154	37.3	l(8.94)
37	rr	0.0324	42.8	e(6.76), r(5.5)
38	zh	0.0023	20.7	y(16), i,r(9.78)
39	t	0.0092	24.7	k(6.63), th(5.57)
40	-	0.0203	94.3	

Table 7.11: Phonetic engine output

Actual transcription (in ASCII) for input word	Transcription given by the phonetic engine (40 models)
SiSirem	sishirem
pathanjali	pathenchali
kaalavastha	kaaalavasta
vismayam	visthmayam

Table 7.11 compare the output of phonetic engine for cetain words with their manual transcription.Inorder to introduce more accuracy and to reduce the likelihood of mismatching the language as Malayalam in a global phonetic engine, we created a new phonetic engine. Here, we chose 26 phoneme classes including silence. Referring to the above 40 classes phonetic engine, grouping was done based on confusable classes.

Most of the nasals are mapped to /n/. Less frequent symbols like /rr/,/f/, /zh/, etc.. are also mapped as per the Table 7.12. We had also increased the data used for PE. Phonetic engine in all the three mode of data were made. The overall accuracy has been increased to 34.14%

The phonetic engine with 26 models were tested for the previous 4 hours of read mode data. It was observed that overall phone recognition correctness was 44.81% and overall phone recognition accuracy was improved to 40.93%. There was 14.59% improvement in phone accuracy due to much reduction in substitution errors.Then phonetic engine with 26 models were tested for all of the transcribed data. It contained 5 hours of read data, and 2.5 hours of conversation and lecture data. 75% of data was divided to training and rest of them to testing, considering gender balance. For read mode, the overall phone recognition correctness is 39.85% and overall phone recognition accuracy is 34.56%. For lecture mode, the overall phone recognition correctness is 39.69% and overall phone recognition accuracy is 32.40%. For conversation mode, the overall phone recognition correctness is 37.13% and overall phone recognition accuracy is 32.45%. It is observed that read mode performs the best among others in recognition accuracy.

We further created a broad phonetic engine with 6 broad classes namely vowel, fricative, approximant, plosive, nasal and silence as per 7.8. Overall accuracy of about 60% was found in read mode.

Table 7.12: Further mapping of 40 phoneme models to 26 phoneme models

Sl. No	26 Phoneme Models	Symbols in 40 phoneme models
1	a	a, aa
2	i	i, ee
3	e	e, E
4	u	u, oo
5	o	o, O
6	k	k
7	g	g
8	ch	ch
9	j	j
10	T	T
11	D	D
12	t	t,th
13	d	d
14	n	n, ng, nj, N, $n\_$
15	p	p
16	b	b
17	m	m
18	L	L, zh
19	l	l
20	r	r, rr
21	y	y
22	v	v
23	sh	S, sh
24	h	h
25	s	s, f
26	-	-

This phonetic engine can be used for audio search applications as explained in next sections.

A Java based GUI was developed for the real time recording of speech followed by its recognition using phonetic engine. Provisions for selection/browsing from already recorded data, playing the recorded data and displaying phoneme outputs in English alphabet using ASCII symbols were incorporated. Screenshot of GUI developed is shown in figure below.

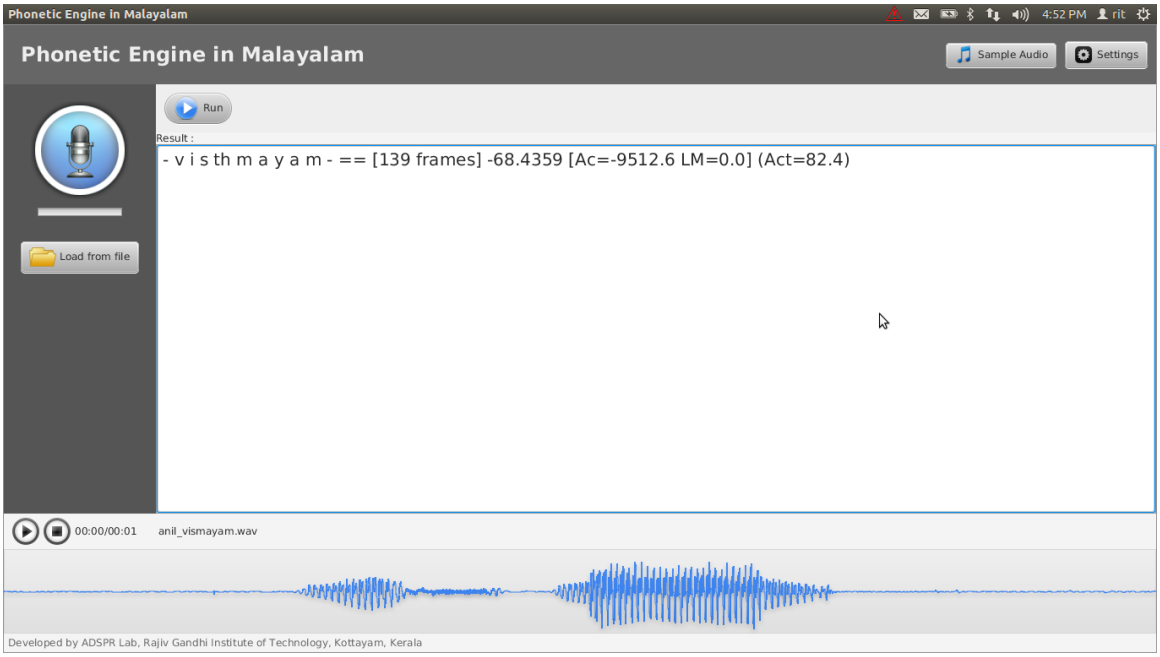


Figure 7.6: GUI developed at RIT for phonetic engine

7.5.2 Development of phonetic engine for Kannada language

The phonetic engine for Kannada was also developed using HTK. HMM models were developed for each mode. HMM developed for each mode ($Eg.Readmode$) is evaluated for the same as well as the other ($Eg.read, lectureandconversation$) modes. The data preparation, preprocessing of data and building of HMMs for Kannada speech corpus is explained in this section.

To build HMM, the data preparation is very important step. This phase consists of labeling of speech signal. The labeled file saved in .lab format is a simple text file with time stamps of speech data whose duration varies from 1 to 8 second. The two and half hour data from each mode is considered, from which 80% of data is used for training i.e. 2 : 00 hours and 20% is used for testing i.e. 30 minutes. Ch-wav command in HTK is used to chop the data files into smaller chunks of about 1 to 8 second and also to obtain their respective transcription. The speech wave files and their transcription with time stamps are considered. IPA symbols were mapped to ASCII equivalents.

Table 7.13: Data and speaker information

Mode	Training duration (Hrs:Mins:secs)	Testing duration (Hrs:Mins:secs)	No.of speakers in training	No.of speakers in testing
Read	01:56:06	00:29:40	11 (5 Male + 6 Female)	2 (1 Male + 1 Female)
Lecture	02:00:31	00:39:14	4 (3 Male + 1 Female)	2 (1 Male + 1 Female)
Conversational	02:04:22	00:27:43	12 (11 Male + 1 Female)	6 (6 Male)

The perceptually similar and least occuring phones are merged together. The merging information is given in Table 7.14. In Kannada the 47 phones are merged together to form 26 phones.

Table 7.14: Phone merging information for Kannada phonetic engine

S.I. No.	Phonetic Unit (in IPA)	Merged Phonetic Unit (in IPA)	Corresponding name (in ASCII)
1	ɔ	a	a
2	ae	a	a
3	aˑ	a	a
4	aː	a	a
5	e	e	e
6	eː	e	e
7	i	i	i
8	iː	i	i
9	o	o	o
10	oː	o	o
11	u	u	u
12	uː	u	u
13	k	k	k
14	g	g	g
15	ɓ	g	g
16	ʒ	j	j
17	ʤ	j	j
18	ʈ	ʈ	T
19	ɖ	ɖ	D
20	t	t	t
21	D	D	d
22	n	n	n
23	p	p	p
24	b	b	b
25	ɱ	ɱ	m
26	j	j	y
27	ɾ	r	r
28	r	r	r
29	l	l	l
30	ʋ	v	v
31	ʃ	ʃ	sh
32	ʂ	ʃ	sh
33	s	s	s
34	h	h	h
35	ɭ	ɭ	L
36	k^h	k	k
37	g	g	g
38	g^h	g	g
39	tS^h	tʃ	ch
40	dZ^h	dʒ	j
41	^{h}	ʈ	T
42	^{h}	ɖ	D
43	t^h	t	t
44	d^h	d	d
45	p^h	p	p
46	b^h	p	p
47	–	–	sil

Initially the HMM is trained and tested for the same mode of data. The performance of HMMs built for read, lecture and conversation modes of Kannada speech are given Table 7.15.

Table 7.15: Training and testing with same modes of data

Training Mode	Testing Mode	Phone Recognition in %	Recognition Accuracy in %
Read	Read	59.36	55.10
Lecture	Lecture	55.80	48.74
Conversation	Conversation	52.34	45.67

From Table 7.15, it can be noticed that the accuracy of read mode data is high compared to other two modes because of proper articulation and it is much closer to the written language. In conversation mode phone recognition accuracy is less compared to other modes because speech consists of filled pauses and non-standard pronunciations.Table 7.16 shows the phone wise classification results for training and testing with same modes of data. From Table 7.16, it is noticed that silence recognized properly in lecture and conversation mode due its frequency of appearance in the training data. The phone L is having least recognition rate. From HMM results in Kannada, L−l, L−r and sh−s are identified as more confusing pairs.

Table 7.16: Phone occurrence rate in training data along with phone wise classification for training and testing with same modes of data

Phone in ASCII	Read Mode		Lecture Mode		Conversation Mode	
	%PO	*%PPR*	*%PO*	*%PRR*	*%PO*	*%PPR*
a	23.72	81.2	22.32	77.1	19.0	19.0
e	5.54	71.5	5.98	76.8	7.06	76.6
i	8.20	77.5	7.22	74.7	7.68	79.8
o	1.15	71.7	2.33	71.4	2.84	73.5
u	6.40	76.6	5.49	73.5	5.42	58.5
k	3.59	76.2	3.56	78.0	3.24	76.7
g	3.10	62.7	3.2	55.3	3.1	60.8
ch	0.70	88.0	0.42	48.6	0.49	52.9
j	0.46	51.5	0.42	68.8	0.57	73.5
T	0.71	47.8	0.87	54.8	2.08	66.1
D	1.16	74.8	1.96	73.8	2.29	70.2
t	4.42	75.7	4.3	71.7	3.75	67.5
d	5.00	67.7	4.85	64.8	4.74	73.1
n	6.46	76.3	8.4	78.2	7.96	70.1
p	1.56	81.0	0.8	60.1	1.37	63.7
b	1.62	68.6	1.92	73.1	1.96	75.1
m	2.68	80.5	2.72	82.7	2.96	79.3
y	2.59	80.2	2.02	67.7	1.76	55.9
r	5.61	86.4	4.49	81.0	5.26	72.4
l	2.91	68.4	2.61	58.3	4.47	67.7
v	3.34	67.4	2.97	60.1	2.26	55.9
sh	1.11	67.7	0.69	83.1	0.62	53.8
s	2.78	85.1	1.82	86.0	2.86	87.6
h	1.24	49.0	1.54	38.1	0.97	49.5
L	1.63	5.2	1.57	17.5	1.1	10.3
sil	2.22	26.4	5.4	74.3	3.84	61.9

$*$*% PO−Phone occurrence in training data*

$*$*% PRR−Phone Recognition Rate*

Table 7.17 shows the HMM results for training and testing with all modes of data. From Table 7.17 it can be noticed that the same mode of training and testing gives comparatively high performance than training and testing with different modes. For example read v/s read gives 55.10% phone recognition accuracy, but read v/s lecture and read v/s conversation gives 34.77% and 40.00% respectively, which are less than lecture v/s lecture and conversation v/s conversation.

From the results of Table 7.17, the experiments are further extended to train HMM by merging different modes of data and test the resulting HMM with all three modes. i.e training HMM by

Table 7.17: Training and Testing with all modes of data

Training Mode	Testing Mode	PR in %	PR Accuracy in %
Read	Read	59.36	55.10
Read	Lecture	42.04	34.77
Read	Conversation	47.43	40.00
Lecture	Read	52.35	46.28
Lecture	Lecture	55.80	48.74
Lecture	Conversation	46.72	37.87
Conversation	Read	48.35	42.99
Conversation	Lecture	47.87	42.29
Conversation	Conversation	52.34	45.67
Conversation	Read	48.35	42.99
Conversation	Lecture	47.87	42.29
Conversation	Conversation	52.34	45.67

read+lecture mode data and testing it with all modes. From Table 7.18, it can be noticed that the mode

Table 7.18: Training by combining different modes of data and testing with three modes

Training Mode	Testing Mode	PR in %	PR Accuracy in %
Read+Lecture	Read	55.93	50.03
Read+Lecture	Lecture	52.35	45.90
Read+Lecture	Conversation	49.11	42.25
Lecture+Conversation	Read	50.65	44.15
Lecture+Conversation	Lecture	53.72	47.93
Lecture+Conversation	Conversation	51.85	44.94
Conversation+Read	Read	57.60	53.69
Conversation+Read	Lecture	45.46	39.61
Conversation+Read	Conversation	50.29	44.24
Read+Lecture+Conversation	Read	57.44	53.41
Read+Lecture+Conversation	Lecture	52.34	46.60
Read+Lecture+Conversation	Conversation	50.65	44.15

which is not present in training set will give less accuracy compared to base line results. For example : Read+Conversation HMM gives 39.61% accuracy for lecture mode data whereas lecture v/s lecture gives 48.74%. The HMM trained by all three modes i.e. HMM for Read+Lecture+Conversation modes give the results which is compatible with the base line results. The Read+Lecture+Conversation mode HMM gives 53.41%, 46.60% and 44.15% accuracies for Read, Lecture and Conversation modes respectively. These results almost compatible with base line results as shown in Table 7.15 i.e. read v/s read , lecture v/s lecture and conversation v/s conversation modes.

7.6 A coarse audio search method using phonetic and prosody labels

It was noticed that the output of the prosody unit will be similar for same keywords occurring at different positions. Audio search helps to locate each and every occurrence of the keyword in the audio database. Here we are suggesting a unique approach for audio search using temporal prosodic pattern. Conventional audio search faces time and computational complexity.

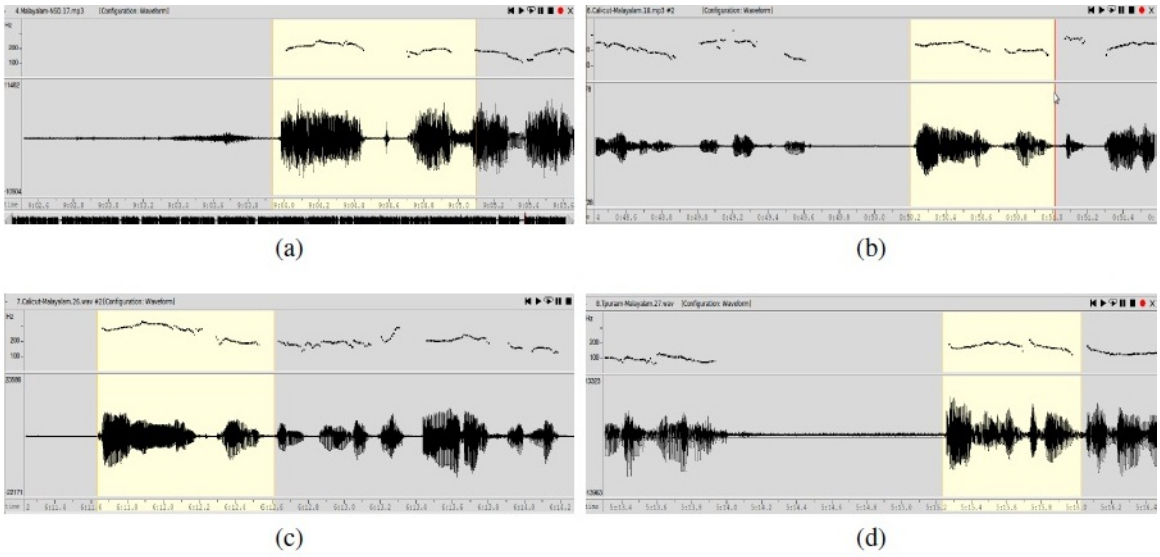


Figure 7.7: (a), (b), (c), and (d), Waveform and pitch contour corresponding to the word *common wealth games* as appeared in AIR news bulletins transmitted by different radio stations

While analyzing the pitch contour of same words by different speakers in different contexts, a similarity is noticed as illustrated in Figure 7.7 and Figure 7.8. There are only smaller variations in pitch trends as we consider same keyword utterance of different speakers. For the words occurring in similar positions, the pith-trends seems to be similar and speaker independent as shown in Figure 7.7 (b), (c), (d) and Figure 7.8(a), (b), (c) . But due to co-articulation, depending on the place of occurrence of the word, the pitch-trend can slightly vary as in Figure 7.7 (a) and Figure 7.8 (d).

This property of pitch trends along with broad phonetic labels may be useful for audio search applications. The methodology proposed is illustrated using block schematic in Figure 7.9. The major blocks in this methodology such as broad phonetic engine, pitch trend labeling, temporal pattern and local alignment are explained below. The broad phonetic engine as explained in earlier section is used here. Methods used for pitch-trend labelling, formation of temporal pattern and local alignment are

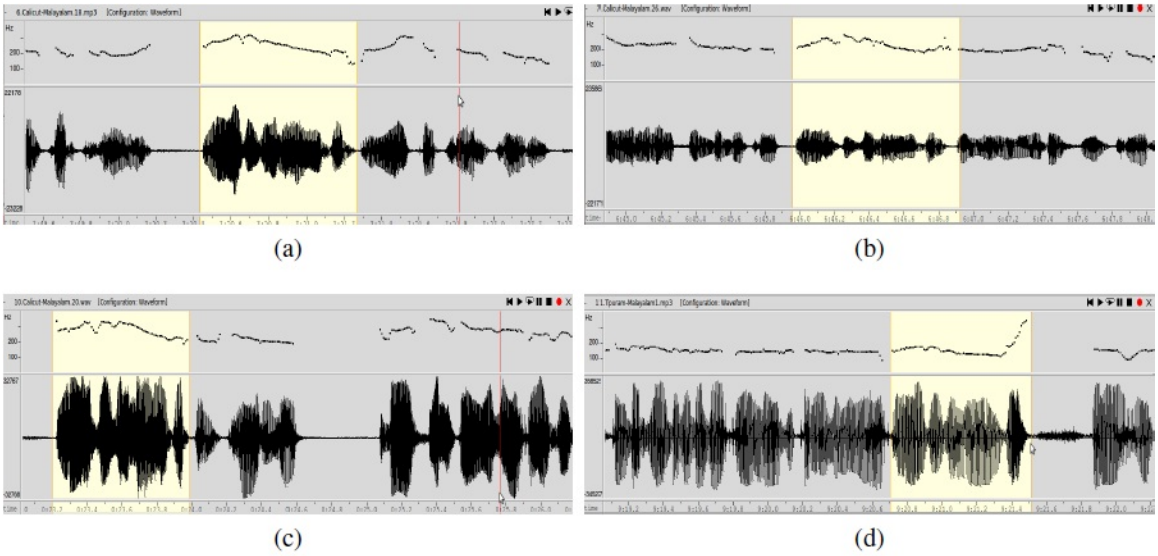


Figure 7.8: (a), (b), (c), and (d), Waveform and pitch contour corresponding to the word *unnathathalayogam* as appeared in AIR news bulletins transmitted by different radio stations

discussed below.

7.6.1 Pitch-Trend Labeling

Trend labeling consists of two parts: (i) Pitch contour modification (smoothing followed by trend removal) and (ii) Transcribing using three labels namely Flat (@), Rise (+), and Fall (∼).

7.6.2 Temporal Pattern

Broad phoneme class label and pitch-trend labels are given as input to this unit. In the temporal pattern block, these two labels are combined and rearranged according to the time stamps. So here the combination of broad phoneme transcription and pitch trend labels is used to obtain a temporal prosodic pattern.

Table 7.19 and 7.20 shows the temporal prosodic pattern for the keyword unnathathalayogam and cheriya perunnal respectively. In the test speech, this keyword is occuring many times. The temporal pattern for several occurrence of the keyword in the database is shown.

To use in audio search applications, the keyword as well as the search database is converted to the temporal prosodic patterns as illustrated in Table 7.21. The temporal prosodic pattern is similar at all the position where the keyword is occurring in the database. So have to measure the similarity measure of the pattern in terms of distance. We can use methods like simple linear distance measure

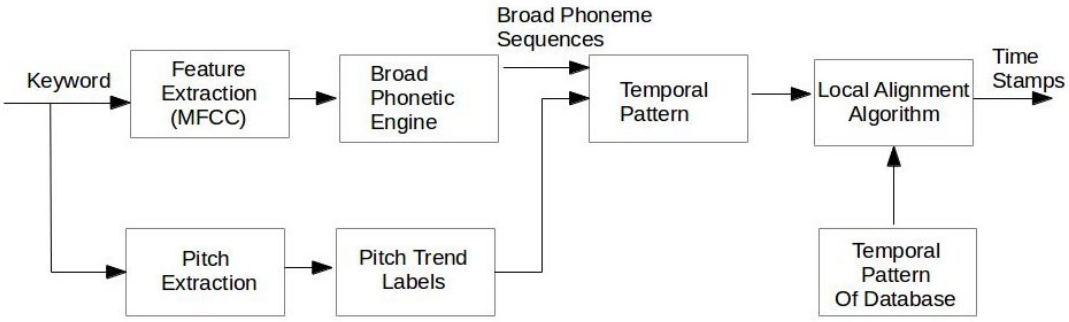


Figure 7.9: Block schematic illustrating the proposed methodology for audio search.

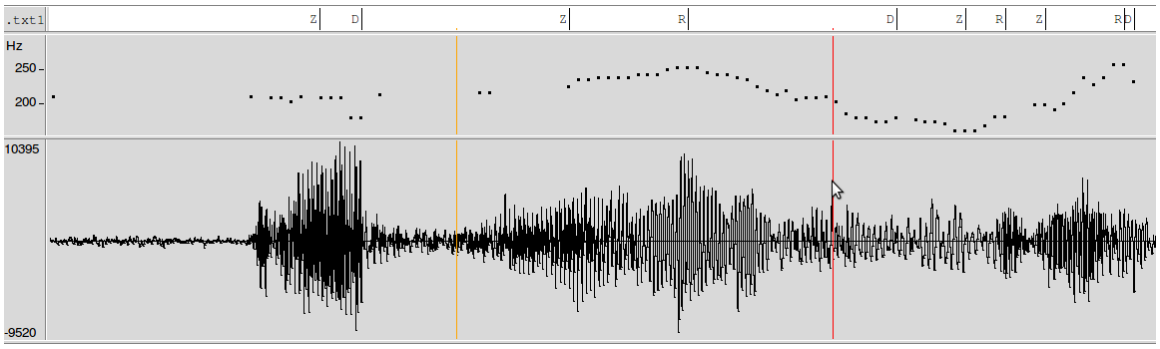


Figure 7.10: Pitch Trend Labeling.

or advanced Dynamic Time Warping (DTW) etc. The position where there is maximum similarity will have the minimum distance measure can be declared as the position of occurrence of the keyword. The Table 7.21 illustrates an example.

7.6.3 Local Alignment

To use in audio search applications here the keyword as well as search database is converted to the temporal prosodic patterns as illustrated in Table 7.19 , 7.20. The temporal prosodic pattern is similar at all the position where the keyword is occurring in the database. It will vary slightly because of the phenomenon termed coarticulation. So we have to calculate the similarity/alignment measure of the patterns in order to locate them.

Sequence alignment is a way of arranging two or more sequences of characters to identify regions

Table 7.19: Prosody model output while testing the word *unnathathalayogam*in different speech signals in the database.

Query word	unnathathalayogam
Broad class label	VNNVPVPVAVAVPVN
Pitch-trend label	@+ ∼ + ∼∼
Occurrence in the test database	Output of prosody model
Occurrence 1	$@VNN + V \sim P@V \sim P@VANV \sim VP@V + N$
Occurrence 2	$@VN + NV \sim +PV \sim P + PV \sim A@AVPV + N$
Occurrence 3	$@VNN + V \sim @P \sim P@VN \sim @A \sim VNV@P + N$

Table 7.20: Prosody model output while testing the word *cheriyaperunnal* in different speech signals in the database.

Query word	cheriyaperunnal
Broad class label	PVAVAV PVAVNAA
Pitch-trend label	@ ∼ + ∼∼
Occurrence in test database	Output of prosody model
Occurrence 1	$@P@AV \sim AV + APV \sim VNA + VA \sim$
Occurrence 2	$@P@V \sim AA + PVNV \sim VNV \sim$
Occurrence 3	$@P@V \sim A + VPV \sim NNVAVA \sim$

of similarity. Aligned sequences are typically represented as rows within a matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns. Sequence alignments are widely used in bioinformatics for arranging the sequences of DNA, RNA, or protein, but are also used for nonbiological sequences, such as those present in natural language or in nancial data. Computational approaches to sequence alignment generally fall into two categories: global alignments and local alignments. Calculating a global alignments a form of global optimization that forces the alignment to span the entire length of all query sequences. By contrast, local alignments identify regions of similarity within long sequences that are often widely divergent overall. Local alignments are often preferable, but can be more difcult to calculate because of the additional challenge of identifying the regions of similarity.

Here we used local alignment method to locate the positions of the keywords with maximum alignment score. This type of search can be considered as the rst stage of audio search, because it locates the positions with maximum alignment. Then we can further rene the search in these selected portions using some other techniques instead of searching in the whole search database. The Table

Table 7.21: An example for illustrating audio search for the keyword *pradhana manthri*

Keyword	pradhana manthri
Broad Class Labels	PAVFVNVNANPAV
Pitch-Trend	DRDRRD
Temporal prosodic pattern of keyword	PA**D**VFVN**R**VNA**D**N**R**PA**R**V**D**
Temporal prosodic pattern of audio database	NAV**D**NPV**D**PVA**R**VNV$\cdots$ PVAFVPAPVN**R**VN**DR**VP**D**A**R**V$\cdots$ PAPN**D**VP**R**VPVNASPANVNVP**RR**V$\cdots$
Duration of audio db	64.680 seconds
No. of broad symbols present in the search file	1088
Actual time-stamps of the positions of keyword in db	0.28, 28.64, 34.09, 37.23, 41.82
Time stamps obtained using alignment (their rank of best alignment)	0.32(4), 28.84(37), 33.71(1), 36.70(21), 41.00(22)
Total no. of best alignment	429

7.21 illustrates examples of audio search.

7.7 Details of training programme conducted

(i) A training programme on "Automatic Speech recognition using Hidden Markov Model", Nov 30-Dec 4 2012 for project staff and M Tech students of RIT and SIT, Tumkur.

(ii) Distinguished lecture on "Advances in speech systems" by prof. Sadaoki Furui,Tokyo Institute of Technology, Japan on Dec 10, 2013.

7.8 Summary & future work

Speech data in three different modes namely, read, lecture mode and conversational speech were collected. Conversational data were collected from different regions of Kerala and Karnataka states. Collected speech data were transcribed using IPA symbols. Transcribed data were analyzed to find out the frequency of occurrence of different symbols. Marking prosody was done in terms of syllabification, break marking and pitch marking using wave surfer. Algorithm for automatic pitch marking and break marking were developed. Automatic syllabification was attempted using prosodic information in terms of pitch and energy contour. Broad phonetic labelling was attempted using signal derived features.

Phonetic engine in Malayalam was developed using 40 HMM phoneme models as well as 26 phoneme models and it's performance was analyzed. HMM based broad phonetic engine was developed which can be used for audio search applications. Phonetic engine in Kannada was developed using 26 HMM phoneme models and it's performance was analyzed. A coarse method for audio search using phonetic and prosodic labels was also attempted. Different methods for integrating prosodic knowledge into the present phonetic engine need to be explored in the future.

8

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT) Gandhinagar

PROGRESS SUMMARY REPORT OF DA-IICT

A. General

A.1	Name of the Project	:	**Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages (Gujarati & Marathi)**
	Our Reference Letter No	:	11(6)/2011-HCC(TDIL) dated 23-12-2011
A.2	Executing Agency	:	DA-IICT, Gandhinagar
A.3	Chief Investigator with Designation	:	Prof. Hemant Patil Associate Professor, DA-IICT, Gandhinagar
	Co-Chief Investigators with Designation	:	Prof. M. V. Joshi Professor, DA-IICT, Gandhinagar
A.4	Project staffs with Qualification	:	Vibha Prajapati (B.A. (Eng.), Data Entry Operator), Maulik Madhavi (Ph.D. Student - Ex Staff member), Shubham Sharma (M.Tech Student - Ex Staff member), Ankur Undhad (M.Tech - Ex Staff member), Bhavik Vacchani (M.Tech - Ex Staff member), Kewal Malde (M.Tech - Ex Staff member), Tarunima Prabhakar (B.Tech - Ex Staff member), Mansi Gokhale (B.Tech - Ex Staff member).
A.5	Total Cost of the Project as approved by DIT	:	
	i) Original	:	50.60 Lakhs
	ii) Revised, if any	:	-
A.6	Date of starting (Indicate date of first sanction)	:	23 December,2011 -
A.7	Date of Completion	:	-
	i) Original	:	-
	ii) Revised, if any	:	-
A.8	Date on which last progress report was Submitted	:	- 28-02-2014

B. Technical

B.1 Works : Progress.(given details in technical report)

- Database collection in three different modes from various dialectal regions in Gujarat and Maharashtra: (in hours:minutes)

Sr No.	Task	Gujarati	Marathi
1.	Read speech	14:00	10:50
2.	Spontaneous speech	12:30	07:20
3.	Lecture speech	07:00	03:00
4.	**Total**	**33:00**	**21:10**

- Statistics of manual labeling work in Gujarat and Maharashtra: (in hours:minutes)

Sr No.	Task	Gujarati	Marathi
1.	Phonetic transcription	11:00	12:30
2.	Syllabification	11:00	12:30
3.	Phone-based labels	11:00	12:30
4.	Pitch and Break marking	01:13	01:19

B.2 Proposed plan-of-work highlighting the action to be taken to achieve the originally proposed targets :

- Report finalization
- Database finalization
- Code delivery
- Finance settlement

C. Financial

Statement of expenditure for 2nd year (in Rupees) (from 1.4.13 to 4.1.14)

(Utilization certificate enclosed) (Figures in Rupees)

Sl. No	**Sanctioned Heads**	**Op. balance as on 1.4.14**	**Funds Received**	**Expenditure incurred (1.4.14 to 31.3.15)**	**Balance as on** 31.3.15
		A		**B**	**A−B**
1.	Capital Equipment	0	0	0	0
2.	Data Collection	267530	0	242697.00	24833.00
3.	Consumable stores	45575	0	45474.33	100.67
4.	Manpower	162933	0	271346.00	-108413.00
5.	Travel	34606	0	35529.00	-923.00
6.	Workshop & Training	129670	0	14830.00	114840.00
7.	Contingency	182672.99	0	186168.88	-3495.89
8.	Coordination & Management	46616	0	0	46616.00
9.	System Integration	0	0	0	0
10.	Over Head (15 %)	0	0	0	0
11.	Interest on Saving a/c	49922	0	0	49922.00
	Total Budget	**919524.99**	**0**	**796045.21**	**123479.78**

D. Project Outcomes

Papers Published:

- Hemant A. Patil, Maulik C. Madhavi, Kewal D. Malde and Bhavik B. Vachhani, "Phonetic transcription of fricatives and plosives for Gujarati and Marathi languages," in *Int. Conf. on Asian Language Process., IALP-2012*, pp. 177–180, Hanoi, Vietnam, 13-15 Nov. 2012.
- Hemant Patil, Maulik Madhavi and Nirav Chhayani, "Person recognition using humming, singing and speech, in *Int. Conf. on Asian Lang. Process. IALP-2012*, Hanoi, Vietnam, pp.149–152, 13-15 Nov. 2012.
- Bhavik B. Vachhani, and Hemant A. Patil, "Use of PLP cepstral features for phonetic segmentation," in *Int. Conf. on Asian Language Process., IALP-2013*, pp. 143–146, Urumqi, China, Aug. 17-19, 2013.
- Hemant A. Patil and Tanvina B. Patel,"Nonlinear prediction of speech signal using Volterra-Wiener series," in *Pros. INTERSPEECH'13*, pp. 1687-1691, Lyon, France, pp. 1687–1691, 25-29 Aug. 2013.
- Hemant A. Patil, Anshu Chittora, and Kewal Dhiraj Malde, "Novel modulation spectrogram based features for obstruent classification," in *Acoustics-2013*, New Delhi, India, Nov., 2013.
- Kewal D. Malde, Bhavik B. Vachhani, Maulik C. Madhavi, Nirav H. Chhayani and Hemant A. Patil, "Development of speech corpora in Gujarati and Marathi for phonetic transcription," *Int. Conf. Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pp.1–6, 25-27 Nov. 2013.
- Anshu Chittora and Hemant A. Patil,"Data collection and corpus design for analysis of nonnal and pathological infant cry," *Int. Conf. Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pp.1–6, 25-27 Nov. 2013.
- Nirav H. Chhayani and Hemant A. Patil,"Development of corpora for person recognition using humming, singing and speech," *Int. Conf. Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pp.1–6, 25-27 Nov. 2013.
- Kewal D. Malde, Anshu Chittora, and Hemant A. Patil, "Classification of fricative using novel modulation spectrogram based features," in *Int. Conf. on Pattern Recognition and Machine Intelligence*, Kolkata, India, 2013, pp. 134–139, Lecture Notes in Computer Science, LNCS, Springer-Verlag, Berlin Heidelberg, Germany, Dec., 2013.
- Yashesh Gaur, Maulik Madhavi and Hemant A. Patil, "Speaker recognition using sparse representation via superimposed features," in *Int. Conf. on Pattern Recognition and Machine Intelligence*, Kolkata,

India, 2013, pp. 140–147, Lecture Notes in Computer Science, LNCS, Springer-Verlag,Berlin Heidelberg, Germany, Dec., 2013.

- Nirmesh J. Shah, Bhavik B. Vachhani, Hardik B. Sailor, Hemant A. Patil, "Effectiveness of PLP-based phonetic segmentation for speech synthesis," *Proc. Int. Conf. Acoust., Speech and Signal Process., ICASSP14*, Florence, Italy, pp. 270–274, May 4-9, 2014.
- Ankur Undhad, Hemant A. Patil and Maulik C. Madhavi, "Exploiting speech source information for vowel landmark detection for low resource language," in *9th Int. Symp. Chinese Spoken Lang. Proc., ISCSLP14*, pp. 546–550, Singapore, 12-14 September 2014.
- Anshu Chittora, Hemant A. Patil, "Classification of pathological infant cries using modulation spectrogram features," in *9th Int. Symp. Chinese Spoken Lang. Proc., ISCSLP14*, pp. 541–545, Singapore, 12-14 September 2014.
- Maulik C. Madhavi, Hemant A. Patil,"Exploiting Variable length Teager Energy Operator in melcepstral features for person recognition from humming," in *9th Int. Symp. Chinese Spoken Lang. Proc., ISCSLP14*, pp. 624–628, Singapore, 12-14 September 2014.
- Anshu Chittora and Hemant A. Patil, "Classification of phonemes using modulation spectrogram based features for Gujarati language," in *Int. Conf. Asian Lang. Process. (IALP)*, pp. 46–49, Kuching, Sarawak, Oct. 20-22, 2014.
- Bhavik Vachhani, Kewal D. Malde, Maulik C. Madhavi and Hemant A. Patil, "A spectral transition measure based Mel cepstral features for obstruent detection," in *Int. Conf. Asian Lang. Process. (IALP)*, pp. 50–53, Kuching, Sarawak, Oct. 20-22, 2014.
- Shubham Sharma, Maulik C. Madhavi, and Hemant A. Patil, "Vocal tract length normalization for vowel recognition in low resource languages," in *Int. Conf. Asian Lang. Process. (IALP)*, pp.54–57, Kuching, Sarawak, Oct. 20-22, 2014.
- Maulik C. Madhavi, Shubham Sharma, and Hemant A. Patil, "Development of language resources for speech application in Gujarati and Marathi," in *Int. Conf. Asian Lang. Process. (IALP)*, pp. 115–118, Kuching, Sarawak, Oct. 20-22, 2014.
- Anshu Chittora, Hemant A. Patil, "Use of glottal inverse filtering for asthma and HIE infant cries classification," in *Int. Conf. Asian Lang. Process. (IALP)*, pp. 158–161, Kuching, Sarawak, Oct. 20-22, 2014.
- Purushotam G. Radadia, Hemant A. Patil,"A cepstral mean subtraction based features for Singer Identification," in *Int. Conf. Asian Lang. Process. (IALP)*, pp. 58–61, Kuching, Sarawak, Oct. 20-22, 2014.

- Shubham Sharma, Maulik C. Madhavi, Hemant A. Patil, "Development of vocal tract length normalized phonetic engine for Gujarati and Marathi languages," in*17th Oriental COCOSDA Conference*, Phuket, Thailand, pp. 1–6, 10-12 September 2014.
- Anshu Chittora, Kewal D. Malde and Hemant A. Patil, "Obstruent classification using modulation spectrogram based features," in *17th Oriental COCOSDA Conference*, pp.1–6, Phuket, Thailand, 10-12 September 2014.
- Shubham Sharma and Hemant A. Patil, "Combining Evidences from Bark scale and Mel scale warped features for VTLN,"in *2nd Int. Conf. on Perception and Machine Intelligence (PerMin)* , C-DAC, Kolkata, pp.133–136, Feb. 26-27, 2015.
- Anshu Chittora, Hemant A. Patil and Kewal D. Malde, "Classification of stop consonants using modulation spectrogram-based features," in *2nd Int. Conf. on Perception and Machine Intelligence (PerMin)*, C-DAC, Kolkata, pp.145–150, Feb. 26-27, 2015.
- Hardik B. Sailor, Maulik C. Madhavi, Hemant A. Patil, "Significance of phase-based features for person recognition using humming," in *2nd Int. Conf. on Perception and Machine Intelligence (PerMin)*, C-DAC, Kolkata, pp.99–103, Feb. 26-27, 2015.
- Purvi Agrawal, Hemant A. Patil,"Fusion of TEO phase with MFCC features for speaker verification," in *2nd Int. Conf. on Perception and Machine Intelligence (PerMin)*, C-DAC, Kolkata, pp. 161–166, Feb. 26-27, 2015.
- Maulik C. Madhavi, Shubham Sharma and Hemant A. Patil, "VTLN using different warping functions for template matching," accepted in *6th Int. Conf. on Pattern Recognition and Machine Intelligence (PReMI)*, Lecture Notes in Computer Science LNCS), Springer, Warsaw, Poland, June 30 - July 3, 2015.
- Anshu Chittora, Hemant A. Patil and Hardik B. Sailor, "Spectro-temporal analysis of HIE and asthma infant cries using auditory spectrogram," accepted in *Int. Conf. on BioSignal Analysis, Processing and System (ICBAPS 2015)*, Kuala Lumpur Malaysia, on 26-28 May 2015.
- Anshu Chittora and Hemant A. Patil, "Analysis of normal and pathological infant cries using bispectrum features derived using HOSVD," accepted in *Int. Conf. on BioSignal Analysis, Processing and System (ICBAPS 2015)*, Kuala Lumpur Malaysia, on 26-28 May 2015.

Papers Submitted:

- Maulik C. Madhavi, Shubham Sharma, and Hemant A. Patil, "A method to alleviate performance degradation in template matching for query in isolation, " manuscript under review for possible publications

in *18th Int. Conf. on Text, Speech and Dialogue (TSD), Lecture Notes in Artificial Intelligence (LNAI)*, Springer, Plzen, Czech Republic, Sept. 14-17, 2015.

- Maulik C. Madhavi, Shubham Sharma, and Hemant A. Patil,"Vocal tract length normalization features for audio search, " manuscript under review for possible publications in *18th Int. Conf. on Text, Speech and Dialogue (TSD), Lecture Notes in Artificial Intelligence (LNAI)*, Springer, Plzen, Czech Republic, Sept. 14-17, 2015.
- Maulik C. Madhavi, Hemant A. Patil and Bhavik B. Vachhani, "Spectral transition measure for detection of obstruents," manuscript under review for possible publications in the *23rd European Signal Processing Conference (EUSIPCO 2015)* , Nice, France, 31st Aug. - 4th Sept., 2015.
- Maulik C. Madhavi, Hemant A. Patil and Nikhil Bhendawade, "Combining evidences from source and system features for spoken keyword detection, manuscript under review for possible publications in the *23rd European Signal Processing Conference (EUSIPCO 2015)* , Nice, France, 31st Aug. - 4th Sept., 2015.

Thesis

- Bhavik Vachhani, "Phonetic Segmentation Unsupervised Approach,"M. Tech Thesis, *Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT)*, July 2013.
- Kewal Malde, "Studies on Transcription, Classification and Detection of Obstruents, " M. Tech Thesis, *Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT)*, Sept. 2013.
- Ankur G. Undhad, "Vowel Landmark Detection for Speech Recognition," M. Tech Thesis, *Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT)*, August 2014.
- Shubham Sharma, "Vocal Tract Length Normalization for Automatic Speech Recognition," M. Tech Thesis, *Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT)*, August 2014.
- Ongoing Ph.D. Thesis: Maulik C. Madhavi in the broad area of audio search.

Appendix 8.1

Detailed Technical Report of Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT) Gandhinagar

Detailed Technical Report of DA-IICT

8.1 Database collection and transcription

DA-IICT team has collected speech data and other relevant meta data in two Indian languages, *viz.*, Gujarati and Marathi. These two languages are spoken mostly in two states of India, i.e., Gujarat and Maharashtra, respectively. The data is recorded in three different modes, *viz.*, read, spontaneous and lecture modes. The data has been collected using portable handy recorder (Zoom H4n) as most of the data was recorded from remote villages and real field environment (i.e., real-life settings). Recording was performed at *44.1 kHz* sampling frequency with 16 bits/sample resolution. For the collection of Gujarati speech data, we have visited several places of Gujarat state to collect voice samples. The places selected includes Gandhinagar (Vavol,Paliyad),Navsari (Moti kakrad, Navsari), Surat, Anand (Umreth),Jamnagar (Vijarkhi, Mota thavariya), Rajkot, Bhavnagar (Chamardi, Bhavnagar) and Kutch (Kera, Anjar). These places covers three dilectal region of Gujarat state, *viz.*,Saurashtra, South Gujarat and North East Gujarat. From Kutch, staff members could manage to collect about *2* Hours of speech data. For the collection of Marathi speech data, we have visited several places of Maharashtra state. The places are mainly includes Ahmedanagar (Kakti), Nanded (Basmath), Latur, Solapur,Sangli (Vibhutvadi), Kolhapur (Ichalkaranji), Pune and Lonavala. These places covers three dilectal region of Maharashtra state. The places are shown by a circle around the surrounding region as in Fig. 8.1 and Fig. 8.2.

Team members would like to thank Prof. B. Yegnanarayana and Prof. Peri Bhaskararao of IIIT-Hyderabad, who shares their immense knowledge on speech technology and phonetic transcription, respectively. Entire DA-IICT Prosody project team attended two workshops on phonetic transcription to learn phonetic transcription using IPA-based symbols. In addition, there were project-related workshops and meetings; which were quite useful. The places for data collection, experiences, observation and various statistics related to phonetic transcription are discussed [3], [4].

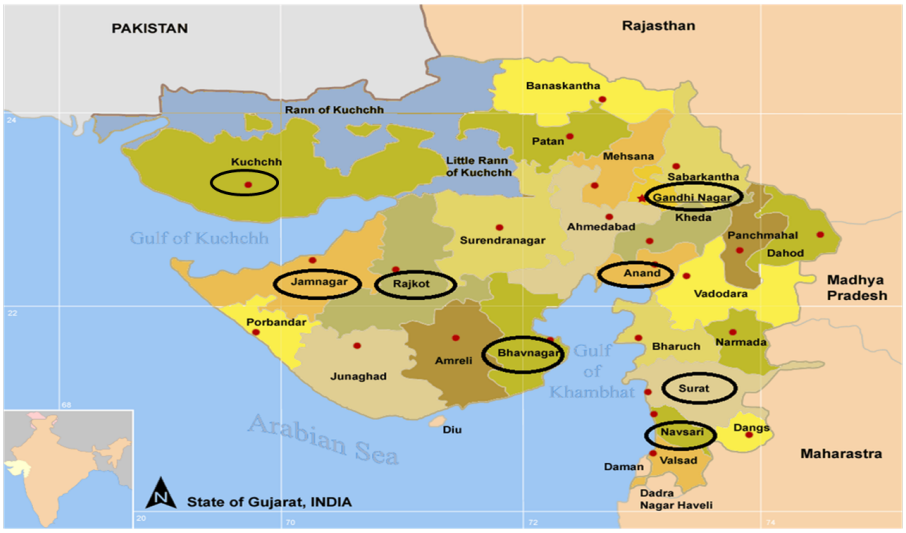


Figure 8.1: Places of Gujarat State. The circles indicate the regions where data has been collected [1].

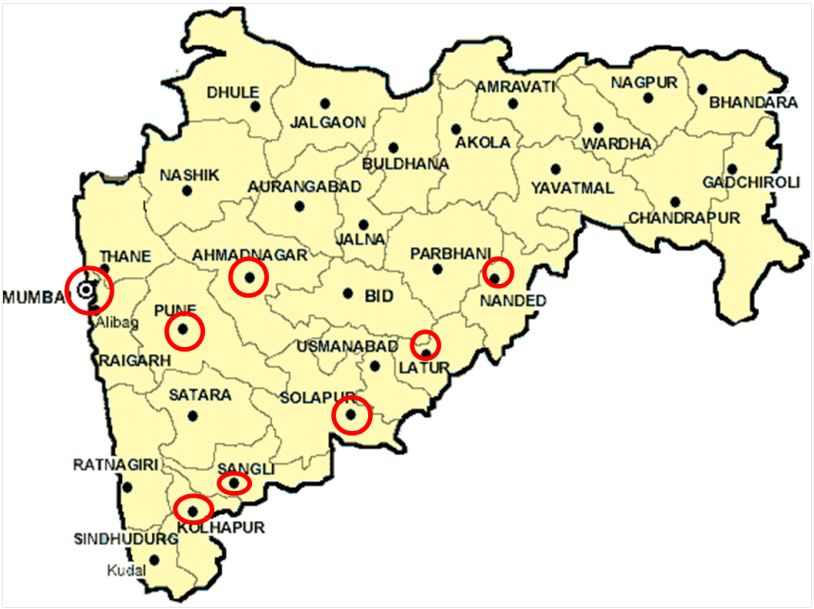


Figure 8.2: A place of Maharashtra State. The circle indicates the region where data has been collected [2].

8.2 Observations

8.2.1 Observation during data collection

Different dialectal zones of Gujarat and Maharashtra are considered during data collection phase. As subjects may not have sufficient knowledge of reading in different kinds of sentences. For example, due to lack of knowledge, subject may speak interrogative type of sentence as a simple declarative. Also, a good reader (i.e., who can read Gujarati and Marathi language fluently) can pronounce each kinds of sentences clearly. As it was observed that reader, who haven't contributed in such data collection may feel shy and nervous while reading. Hence, read mode speech has got less *prosodic* information. This is the problem with speech data collection in read mode. Hence, in order to capture the prosody information, we need to collect the data in spontaneous mode as well.

8.2.2 Observation while transcribing the speech

The speech transcription activities typically involves, tagging speech events using IPA-based phonetic labels at different levels, syllabification task, marking prosodic breaks and variation via break and pitch marking, respectively. Full time and part time consultants are given the task of phonetic transcription and syllabification. Since consultants are graduate at various discipline, they are given prior training for phonetic transcription. The followings are the observations found while transcribing the speech signal.

- Many times listener finds overlap across two phonetic symbols.
- Due to ambiguity between aspirated plosive and fricative sounds, transcriber often confuses [3].
- Human perception of phonetic symbols at different-level is different. It means that a person may not recognize the same phonetic symbol at word or syllable-level than at sentence-level.
- Any two transcribers may not identify the exact the same phone and word boundaries.
- Diacritic marks are very error prone in terms of agreement between two transcribers.
- In a lecture mode, speech subject tries to prolong the vowels in order to create interest among listeners. (Here, children of primary school are the listeners mostly.)

- Transcribers also confuse between fricatives and affricates. This might be due to the improper analysis window selection before listening. It was found earlier from the experiments that frication noise duration is the important cue which distinguish fricatives from affricates [5]. For example, in a syllable (in /CV/ form) containing fricatives in /C/ consonant position is considered partly along with following vowel, a listener may perceive it as affricates with the same vowel.

- In both the languages, diphthongs and associated vowels may be perceived as two distinct vowels. Hence, transcriber may mark as two different syllables instead of vowel.

- Furthermore, transcribers face few software handling difficulties such as saving and editing the labels in wavesurfer. However, transcribers overcome such difficulties as time progresses.

- Different observations are also found while pitch marking. Pitch variation for the same sentence for different speakers may be same or different.

- In a declarative sentence, it is often found that pitch marking is high initially and low at the end.

- Pitch variation in the sentence need not be gradual. Sometimes it changes in between two prosodic breaks.

- Prosodic breaks (i.e., break marking) are not physical pause in between the speech signal. Fluency, hesitation while speaking, also play an important role in break marking.

From a phonetic transcription, a syllable cluster is formed. A syllable is a linguistic abstraction of speech, which accommodates one syllable peak as a vowel. Based on the formation, syllables can be considered to have *7* different types, i.e., *V*, *CV*, *VC*, *CVC*, *C*V*, *VC**, *C*VC**, where *V*, *C* and *C** stand for any vowel, consonant and more than one consonant units, respectively. If *C* is attached before *V*, *C* is called onset part of a syllable and if *C* is attached after *V*, *C* is called code of a syllable. In general, a syllable can have one or many onsets and/or coda. If it contains multiple onsets or coda, syllable is said to be a complex syllable. Table 8.1 and Table 8.2 show the statistics of different kinds of syllable structures observed in both the languages in the *three* different recording modes.

From Table 8.1 and Table 8.2, it can be inferred that most of the syllables are of *CV* types, which is almost more than *50* % of entire syllable coverage. It is indeed correct since writing script in

Table 8.1: Statistics of different types of syllables obtained in manual syllabification (numbers indicate % coverage) for Gujarati language.

Gujarati	V	CV	VC	CVC	C*V	VC*	C*VC*
R	6.41	**59.52**	2.02	29.01	2.77	0.22	0.05
L	10.06	**58.23**	3.06	25.58	3	0.07	0.01
C	9.08	**61.98**	2.08	24.55	2.14	0.15	0.03

Table 8.2: Statistics of different types of syllables obtained in manual syllabification (numbers indicate % coverage) for Marathi language.

Marathi	V	CV	VC	CVC	C*V	VC*	C*VC*
R	6.79	**56.75**	2.02	30.53	3.84	0.05	0.02
L	5.15	**62.27**	1.31	26.65	4.49	0.02	0.01
C	8.41	**62.85**	1.62	23.36	3.7	0.04	0.01

Indian language allows decomposing the basic graphemes into consonants followed by vowel, forming $/CV/$ type of syllable units. Hence, $/CV/$ type syllables are general form of syllable units in Indian languages and hence they are expected to be more frequent. In addition, it can also be observed that the number of syllables formed by consonants in code position is very less, this might be due to difficulty in terms of pronunciation. Same for complex syllable, it is very less dominant in syllable coverage.

8.3 Development of Phonetic Engine (PE)

Phonetic Engine (PE) is the tool to convert speech information into some appropriate symbolic representation. Here, as a part of project objectives, project members follow IPA-based symbols as a phonetic symbols. For this purpose, segmentation and labelling-based approach can be applied. The segmentation is to be performed onto the continuous speech data, so that speech is decomposed into small speech sound units like phoneme and at appropriate label assignment (to a phoneme) is done. While testing phase, phoneme units of segmented speech is classified and relevant phone label is assigned. Here, exploiting the knowledge of speech production mechanism and finding the specific events in the continuous speech would definitely be useful. The events could be voicing detection, place of articulation, manner of articulation and so on. Based on the detection of such events appropriate phoneme label is assigned. So far DAIICT team has been involved in the segmentation and isolated phone recognition task. The segmentation at phone-level and syllable-level could be performed using

Spectral Transition Measure (STM) [6]. It was observed that STM can indeed be useful representation for phonetic segmentation application [7]. Since phonetic transcription is very closely related to the speech signal, one has to extract the information from speech signal itself. The approach used here is based on segmentation and decoding.

The speech segmentation is very crucial component in the phonetic engine design. There are few issues in segmentation task. One of the issues is as to which level segmentation needs to be performed (phone *vs.* syllable). The second is number of a segment which is due to unsupervised nature of the segmentation. The segmentation is performed using Spectral Transition Measure(STM). To compute STM contour cepstral information is computed from the speech signal. Then STM at time i^{th}, instance STM_i is defined as

$$STM_i = \frac{1}{K}\sum_{i=1}^{K} a_i^2, \tag{8.1}$$

where $a_i = \frac{\sum_{k=-I}^{I} kc(i+k)}{\sum_{k=-I}^{I} k^2}$, and $c(i)$ is the i^{th} order cepstral coefficients [6]. More details of phone segmentation using STM is discussed in Section 8.4.1.

A Gaussian Mixture Model (GMM) of *4* Gaussian components is trained for an individual isolated phone. The speech data is converted into its Mel Frequency Cepstral Coefficients (MFCC) representation having window duration *16* ms and shift of *4* ms. GMM decoding is performed at frame-level and which is fluctuating if a hard decision made upon a frame-level, so GMM decoding result is averaged out over a segment (which contains multiple frames) to estimate the phonetic sound unit. For training task, speech signal is segmented manually at phone-level. The output of a phonetic engine is displayed in Fig. 8.3 and Fig. 8.4. There are few observations are found from above design.

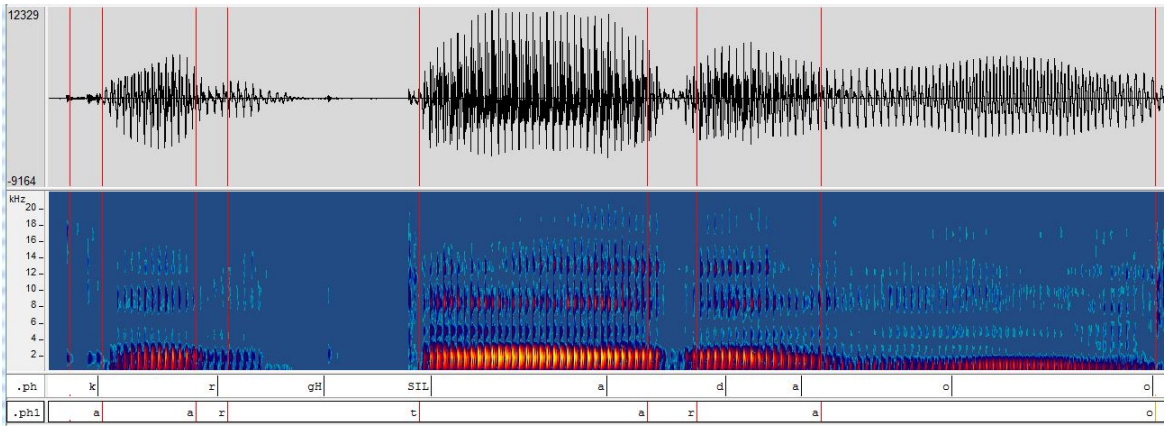


Figure 8.3: A time-domain waveform, spectrogram, output of Gujarati phonetic engine and expected outcome, i.e., manual symbols.

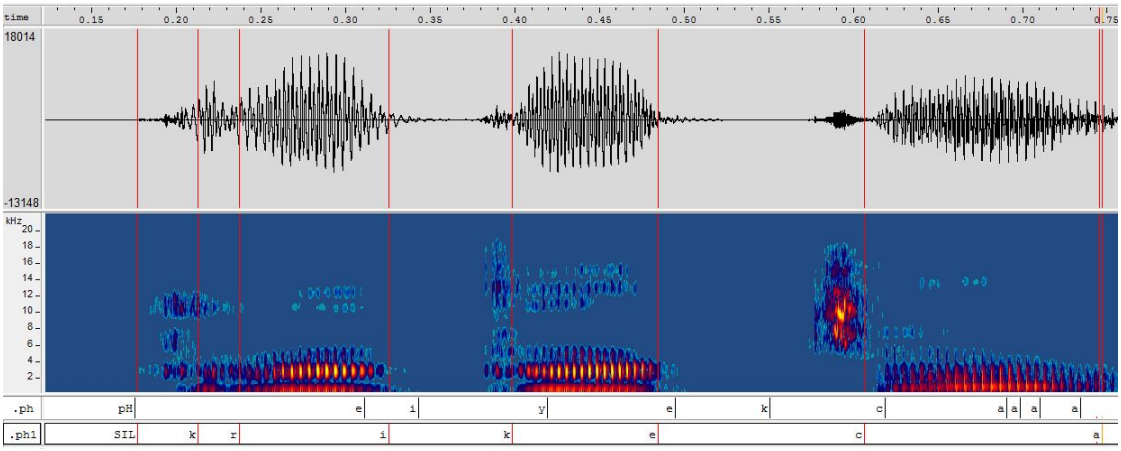


Figure 8.4: A time-domain waveform, spectrogram, output of Marathi phonetic engine and expected outcome, i.e., manual symbols.

- Segmentation plays very important role. Many times, it is found that vowels are broken in multiple segments.
- This might not affect the system much. However, if vowels and consonants are not separated during segmentation the performance might be biased based on the averaged log- likelihood scores.
- Here, segmentation is *unsupervised* and decoding is *supervised*. Manual labels have to be correct which are used in training.
- Scoring in GMM framework is at frame-level, so the closure duration before the unvoiced plosive are treated as silence for frame. Hence, again segmentation has to be improved.

8.3.1 HMM-based PE:

The actual phonetic engine is based on statistical model (HMM). Two different sets of feature vectors, *viz.*, MFCC and PLP are used in the development of PE. Features are extracted over window duration of *25 ms* and shift of *10 ms*. For speech samples of *16 kHz* sampling frequency, each frame consists of *400* samples. Each feature vector consists of *13* static coefficients (C_0 to C_{12}). Δ and acceleration (i.e., $\Delta - \Delta$) coefficients are calculated. This makes the feature vector of dimension *39* (*13* static + *13* delta (Δ)+ *13* acceleration ($\Delta - \Delta$) coefficients). For feature extraction, *28* Mel spaced subband filters are used in MFCC and the order of LPCs is taken as *12* in PLP feature extraction.

PE system is designed based on train-test philosophy of phonetic sound units. Phonetic units that are manually transcribed by the transcribers are used to train HMMs for each phonetic unit using HTK (HMM Toolkit) [8]. Since the speech signal is not aligned with respect to every phonetic symbol, flat start based approach is employed. Five state HMMs which include two nonemitting and three emitting states with single Gaussian model per state are initialized. HMM embedded re-estimation is performed several times. Finally, the test data is decoded into single phonetic string. No phone language model (LM) is used here, since it is expected that consecutive phone sequence might not capture effective information which is derived from manual phonetic transcription. Similar design procedure is used to develop PEs for both the languages, *viz.*, Gujarati and Marathi and for all recording modes. PE is designed using the two kinds of features, *viz.*, MFCC and PLP and their performance in the three modes of speech is shown in Table 8.3. Performance is measured in terms of percentage number of correctly recognized phonetic symbols. It is given by

$$\% \text{ Correctly detected units} = \frac{H}{N} \times 100, \tag{8.2}$$

where H is the number of correctly recognized phonetic units and N is the total number of phonetic units in the transcription.

8.3.1.1 Result:

From Table 8.3, the performance for read speech is observed to be better (in both Gujarati and Marathi databases) as compared to spontaneous speech and lecture speech. This may be due to the fact that read speech has least prosodic variations whereas lecture speech has higher variations in intonation (and thus speech prosody in general). In read speech, the speakers are constrained by the given fixed text material and hence, there are less prosodic variations which is not the case in spontaneous and lecture speech. On comparing the performance of the two features (such as MFCC and PLP), it is observed that PLP features have better performance over MFCC features for most of the cases. This is due to the fact that PLP features are derived from perceptual properties and are speaker-invariant.

The following observations are made from the confusion matrix for read speech of Gujarati database using MFCC. Major classification mistakes happen with aspirated and non-aspirated forms of consonants. Most of the aspirated consonants are observed to be misclassified to their non- aspirated

Table 8.3: Classification (in %) correct detection of MFCC and PLP in classification of phonetic units for Gujarati and Marathi.

Features	G-R	G-C	G-L	M-R	M-C	M-L
MFCC	67.11	62.37	59.84	59.19	49.81	39.64
PLP	66.89	62.75	60.18	60.36	48.82	41.76

versions. For example, most confusing aspirated consonants are [b] - [bH], [c] - [cH], [d] - [dH], [g] - [gH], [k] - [kH], [p] - [pH] and [t] - [tH]. The basic difference between the aspirated and non-aspirated consonants is that in aspirated ones, an aspiration occurs simultaneously with the voicing. The reason for non-aspirated consonants being detected as aspirated ones might be the presence of some noise followed by the consonant that is being detected as *aspiration*. On the other hand, the aspiration part of aspirated consonants may be missed leading to misclassification as non-aspirated. In addition, as most of the Indian languages have phones followed by schwa (i.e., [a]), this results in confusion for transcribers as to whether to put [a] or not and human errors take place. It is observed from confusion matrix that schwa is confused with almost all of the phonemes and there have been a large number of insertions and deletions. This type of misclassification can be reduced to a certain extent by improving and making precise transcriptions. For very small occurrences, silence is detected as plosives such as [T], [t], [k], [p], etc. Presence of bursts might be detected due to the presence of unavoidable noise in the real field environment. It is found that most of the times, vowel gets confused with vowels, such as, [A] gets confused with [a], [e] and [o]; [e] gets confused with [i] and vice-versa; [o] gets confused with [a] and [u]. In addition, plosive consonants get confused with plosive consonants. For example, [t] gets confused with [T], [d], [k] and [p]; [d] gets confused with [D], [b] and [g]; [b] gets confused with [d] and [g]. This is because of the short durations of plosives which are not easily captured even though delta and acceleration (i.e., delta-delta) coefficients are used to capture dynamics of vocal tract. Fricatives get confused with other fricatives, such as [z] gets confused with [j] and [s] gets confused with [z] and nasals get confused with other nasals, such as [m] gets confused with [n]. Another observation is misclassification of fricative [s] as aspirated consonants like [cH] and [pH]. Similar analysis is observed across different phonetic representation.

8.3.2 Syllabification

A speech syllable contains vowel within it. A vowel is found to have relatively higher short-term energy (STE) than consonant units. STE-based information can be an important candidate for syllable-based clustering. The speech segmentation at syllable-level is performed using minimum-phase group delay approach. The detailed description of the algorithm is given in [9]. For boundary detection, we use lowpass version of speech signal having cut-off frequency of *500* Hz. Window Scale Factor (WSF) is chosen *10* and γ is taken as *0.01* as suggested in [9]. The lowpass version of signal is taken because of the fact that vowels and syllable energy is more concentrated at low frequency zone (Typical speech vocal source vibrating frequency, i.e., F_0 is typically less than *500* Hz). Here, data points are tagged by $< L > - < R\_m >$, where $< L >$ can be Gujarati and Marathi, represented by G and M respectively; and $< R\_m >$ can be read, lecture and conversation, represented by R, L, and C respectively.

8.3.2.1 Performance Evaluation

Performance is evaluated for different agreement windows, which change w.r.t. adjacent syllable duration. The evaluation metrics are % detection rate (% DR) within syllable agreement duration and % over segmentation within agreement (% OSWA) and % over segmentation outside agreement window (% OSOA). % DR should be high and over segmentation should be low. x % agreement interval for i^{th} segment, is defined as:

$$\zeta_i - \frac{x}{100}\left(\zeta_i - \zeta_{i-1}\right) \leq \epsilon_i \leq \zeta_i + \frac{x}{100}\left(\zeta_{i+1} - \zeta_i\right), \tag{8.3}$$

where ζ_i's are the manually marked syllable boundaries. Formally, evaluation metrics are defined based on the position of hypothetical boundary (HyB) and agreement interval ($AgInt$) as follows:

$$\%DR = \frac{\#\text{Times } HyB \text{ fall within } AgInt}{\#\text{Total referece boundaries}} \times 100\ \%, \tag{8.4}$$

$$\%OSOA = \frac{\#\text{Times } HyB \text{ fall outside } AgInt}{\#\text{Total } HyB} \times 100\ \%, \tag{8.5}$$

$$\%OSWA = \frac{\#\text{Times } HyB \text{ fall inside } AgInt}{\#\text{Total } HyB} \times 100\ \%, \tag{8.6}$$

8.3.2.2 Database

The corpus is prepared for two languages, *viz.*, Gujarati and Marathi and three different recording modes, *viz.*, read, conversation, and lecture. The sentences are selected such that they contain at least *10* syllables. The statistics of number of sentences used and duration is shown in Table 8.4.

Table 8.4: Statistics of data used in syllabification task [10].

Gujarati	**Read**	**Conversation**	**Lecture**
Sentences	2331	1619	1371
Duration (minutes)	295	200	150
Marathi	**Read**	**Conversation**	**Lecture**
Sentences	3128	899	848
Duration (minutes)	373	90	122

8.3.2.3 Experimental Results

The results obtained from the segmentation task is shown in Fig. 8.5, Fig. 8.6 and Fig. 8.7 for Marathi and Gujarati. The value of %AgInt varies from *5* to *50* % in the steps of *5* %. It can be observed that as % AgInt increases, % detection rate increase and at the same time, over segmentation increases. It means that there is a trade-off between detection rate and over segmentation. In addition, it can be observed that the performance of read mode is better than conversation mode and lecture mode. The % *OSWA* increases exponentially w.r.t. % *AgInt*, as it might be generalized Poisson distribution as it counts number of events within interval. % *OSOA* decreases linearly w.r.t. % *AgInt*. % DR also increases w.r.t. % *AgInt*.

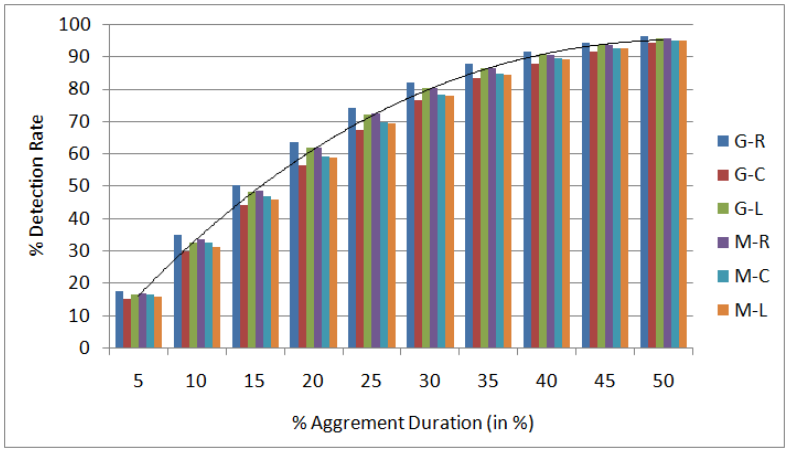


Figure 8.5: Performance of % Detection rate in automatic syllabification task [10].

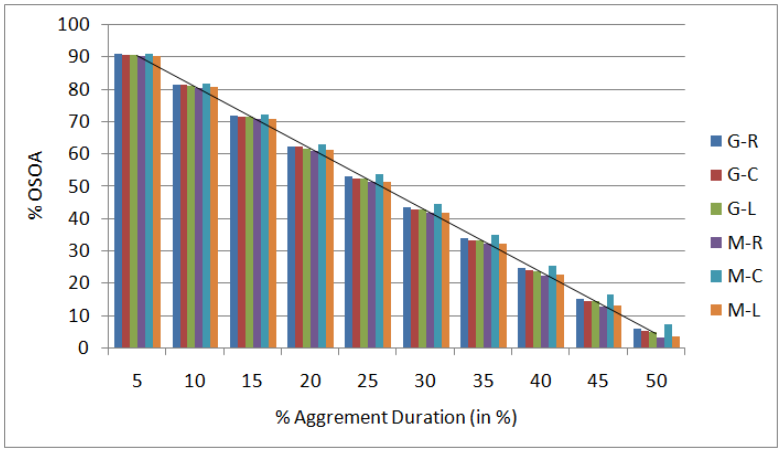


Figure 8.6: Performance of Over segmentation Outside Agreement automatic syllabification [10].

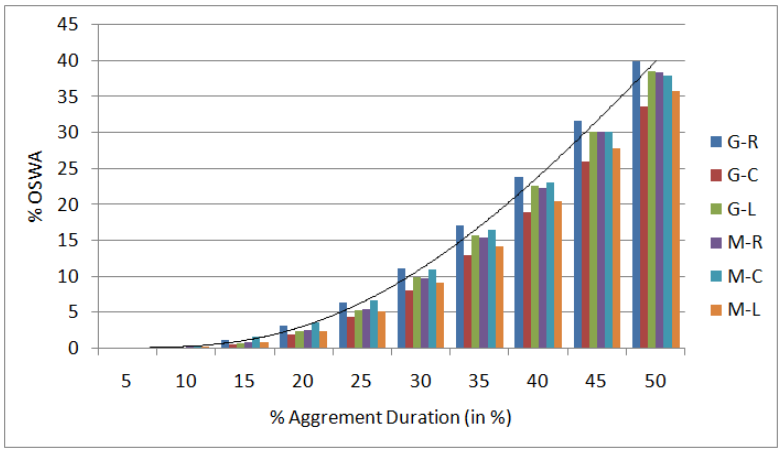


Figure 8.7: Performance of Over Segmentation Within Agreement automatic syllabification [10].

8.3.3 Pitch Marking

Pitch marks are one of the important prosodic events for the implementation of phonetically guided search engine. Pitch is a perceptual attribute of sound which may be physically related with the rate of vibration of the vocal folds. Manual marking of pitch levels is a cumbersome task and varies with individual. Hence, automation of pitch level marking is important. For our implementation, we have assigned four levels of pitch, *viz.*, very low (VL), low (L), high (H) and very high (VH). These levels are determined by the relative increase or decrease of the F_0 contour points [11].

F_0 contour is calculated using the zero frequency filter algorithm. The first step is to median filter the F_0 contour so as to smoothen the contour. Pitch marking is done mostly at the syllable boundaries as the pitch variations are mostly observed around syllable boundaries. Hence, segmentation of the

speech is done at syllable-level. Average of the pitch contour is taken over a period of $5sec$ as a reference pitch ($Fref$) with respect to which pitch levels are marked as VL, L, H and VH. Following algorithm is adopted for automatic marking of pitch-levels:

(i) F_0 contour is found for each utterance using *0-Hz* resonator [12].

(ii) F_0 is interpolated using spline interpolation to find pitch frequencies at each sample point to facilitate the comparison with ground truth.

(iii) Mean F_0 is computed for each utterance known as reference pitch ($Fref$).

(iv) Each sample point is assigned a pitch-level (VL / L / H / VH) according to the following conditions:

a. If $F < 0.5Fref$, pitch-level = VL

b. If $0.5Fref \leq F < Fref$, pitch-level = L

c. If $Fref \leq F < 1.5Fref$, pitch-level = H

d. If $F \geq 1.5Fref$, pitch-level =VH.

(v) To find the accuracy, pitch-level at ground truth is compared with pitch-level automatically determined.

The performance of syllabification system is evaluated using % Accuracy, which is formally defined as,

$$\%Accuracy = \frac{\#\text{ Correctly determined pitch levels}}{\#\text{ Total pitch levels}} \times 100\%. \tag{8.7}$$

Further details of the algorithm of automatic pitch marking are discussed in [11]. The performance of pitch marking along with statistics is shown in Table 8.5. It can be observed that the performance

Table 8.5: Statistics of data used in pitch marking and performance of pitch mark detection system.

Gujarati	**Read**	**Lecture**	**Spontaneous**
Duration	48	28	19
Total pitch marks	6354	5149	4472
Accuracy	63.33%	54.08%	46.82%
Marathi	**Read**	**Lecture**	**Spontaneous**
Duration	56	20	20
Total pitch marks	16619	6059	6830
Accuracy	58.89%	41.98%	43.32%

of read mode is better than lecture and spontaneous mode. The reason could be that the arithmetic used in *step 4*, is not generalized for spontaneous and lecture mode speech as these modes of speech contain more variation in F_0 pattern than read speech. In addition to this, prosodic breaks can be detected using discontinuity in F_0 pattern and STE profile.

8.3.4 Break Marking

Break marking problem is posed as speech detection problem, i.e., to identify the marking such as B_0, B_1 and B_2, speech and non-speech regions are identified. One approach could be use of classification of frames as speech or non-speech based on average short-time energy (STE). However, it might be affected from noisy. Here, we invoke speech signal processing knowledge in order to detect breaks in continuous speech signal. There is an role of formant frequency in the production of most of the speech sound units and hence, it can be better candidate for speech activity and break detection. Conventional Short-Time Fourier Transform (STFT)-based formant peaks are not having large dynamic variations as well as formant peaks are blunted for few cases as suggested in [13]. Here, in this work, formant-like information extracted from modified group delay-based approach is used. The details of algorithm is as follows.

(i) Compute the most dominant peaks from modified group-delay extracted for every frames (Frame duration is *20 ms* and frame shift is *10 ms*).

(ii) Smooth out the contour obtained from earlier step using *200 ms* $\sim$ *20 points* Hamming window.

(iii) Apply hard thresholding over every speech utterance. The threshold value is set as $\theta = \mu - \sigma$.

(iv) Smooth out the contour obtained which takes binary values from earlier step using *150 ms* $\sim$ *15 points* Hamming window. This is used since human perceives no break between *150 ms*.

(v) All the segments where contour estimated from earlier step falls from or to *1* are considered to be hypothetical breaks.

The performance of break marking is evaluated in terms of correct detection and false detection. Formally speaking,

$$\% \text{ Correct Detection} = \frac{\#\text{ times refercence break marks fall within detected segment}}{\#\text{ reference break marks}} \tag{8.8}$$

$$\% \text{ False Detection} = \frac{\#\text{ segments does not associated with refercence break marks}}{\#\text{ segments}} \tag{8.9}$$

The performance is listed in Table 8.6.

Table 8.6: Performance of break marking task.

Gujarati	**% Correct Detection**	**% False Detection**
Read	79.07	59
Conversation	63.18	51.02
Lecture	91.05	35.94
Marathi	**% Correct Detection**	**% False Detection**
Read	43.97	30.3
Conversation	67.43	46.29
Lecture	77.49	27.07

Table 8.6, shows that lecture mode speech is effectively marked at break levels using described algorithm. This might be clear pauses left by subject, who are mainly primary school teachers. Performance of read and conversation less which might be the presence of filler and hesitation while speaking at promptly. This refers to use other information in order to detect the breaks.

8.3.5 Search Engine

The search engine task is to search audio segment within audio data. While matching two different speech-patterns, in practice, few part of one pattern is stretched while remaining is not. This issue can be solved by Dynamic Time Warping (DTW) algorithm. Few constraints may employ to match two patterns dynamically. However, linguistic content of query and data is different. In general, duration of audio query is quite smaller than the audio data. Hypothetical segments are prepared onto test utterance of the size slightly higher or smaller than query length. The segment window is shifted such that the shifted segment has some overlap with the previous segment. The segmental DTW is the variant of DTW algorithm where the (last) ending points is not known [14]. Fig. 8.8 shows an illustration of application of segmental DTW for audio search task.

In addition, test pattern is not speaker-invariant, so speaker-invariant speech representation is important in the audio search task. Few techniques such as posteriorgram representation and Vocal Tract Length Normalization (VTLN) could be exploited.

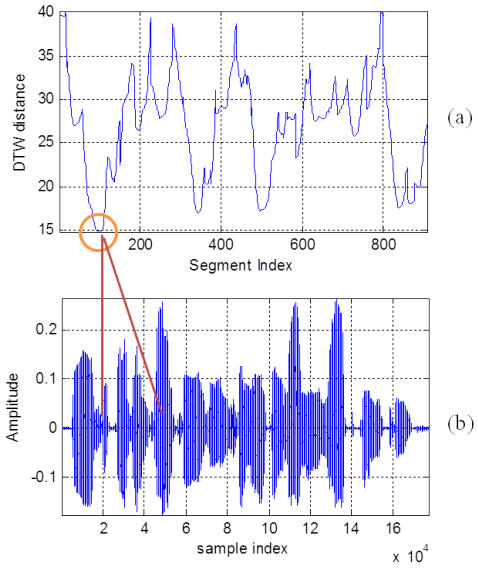


Figure 8.8: (a) DTW- distance for various segments (b) audio data, Gujarati sentence, 'mane gujaraati bolataa aavaDe che'.

8.4 Research activities in the context of project outcomes

DA-IICT Prosody team has been involved in following project related technical and research activities.

(i) Phonetic segmentation of speech signal.

(ii) Broad phonetic classification.

(iii) Keyword search in spoken database.

(iv) Classification of unvoiced fricatives.

(v) Vowel landmark detection in speech.

(vi) Vocal Tract Length Normalization (VTLN) for ASR task.

These are described in brief in next sub-Section.

8.4.1 Phonetic Segmentation

The phone is a smallest acoustic unit of pronunciation. It is the link between the speech signal which is continuous and the phoneme which is just discrete. Acoustic-phonetic segmentation is a task

of detecting the boundaries of phones in a speech signal and labelling those speech segments with phones [15]. Speech segmentation is a process of identifying the boundaries between sound units such as words, syllables or phones in a spoken utterance. Identification of boundaries at phone level is a difficult problem due to the phenomenon of co-articulation of speech sounds, whereas one sound may be modified in various ways by the adjacent sounds due to which the sounds may split or even disappear or perceived as a different sound unit. This phenomenon may happen between adjacent words just as easily as within a single word [16].

A signal processing-based approaches can be used for phonetic boundary detection. These approaches generally use the combination of signal processing techniques and peak-picking methods to perform the task of acoustic-phonetic boundary detection. Mostly, this approach falls in unsupervised and unconstrained category as they do not require any manually labelled speech data. Primarily spectral-based information is used in this approach. One of the approaches is based on *Spectral Transition Measure* (STM). Dusan and Rabiner, uses a spectral transition measure (STM) to capture the spectral rate of change in time [7]. To capture spectral behaviour, *10*-dimensional Mel Frequency Cepstral Coefficients (MFCC) is used over *10* ms frame duration. It expected that spectral rate of change usually displays peaks at the transition between phones; such a measure is used to detect the phone boundaries. Once STM is obtained at every frame, we obtain STM-based contours. Next step is to detect the boundaries using peak picking method and remove spurious peaks using post processing methods [7]- [6]. An experiment was performed on Gujarati, Marathi and an utterance taken from TIMIT database. The results obtained are shown in Table 8.7 and Figs. 8.9- 8.11. Some of the observations from the results obtained are as follows.

- The STM contour is high at the start of frication.
- We obtained better results for plosive and fricative sounds.
- For diphthongs, STM peaks are found approximately at center which is *ambiguous*.

We used MFCC and CFCC features to capture the spectral changes. We proposed a fusion strategy which is F1 and F2 , which uses evidences from MFCC and CFCC features. The results for proposed feature are shown in Table 8.8.

One of the team member of DA-IICT prosody team, *viz.*, Bhavik Vachhani has completed his M.Tech thesis in the area of *phonetic segmentation*.

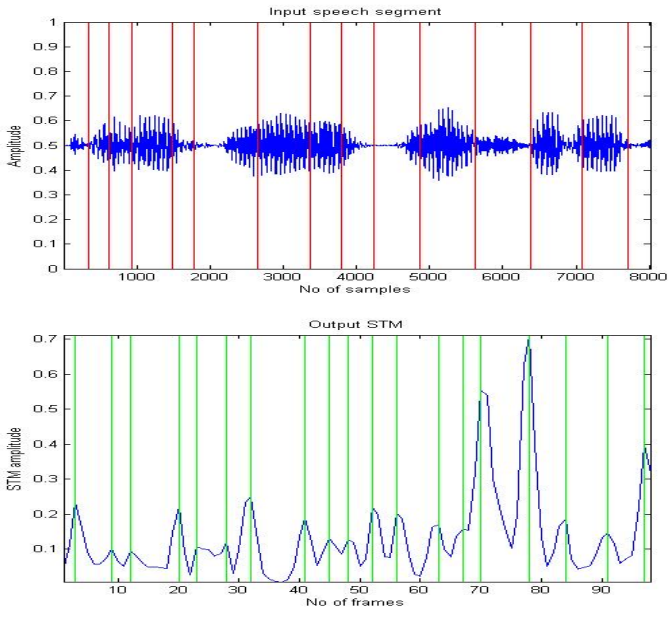


Figure 8.9: (a) manual and (b) automatic segmented phonetic boundary of Gujarati sentence, "kheDuutanaa ruupamaaM sharu".

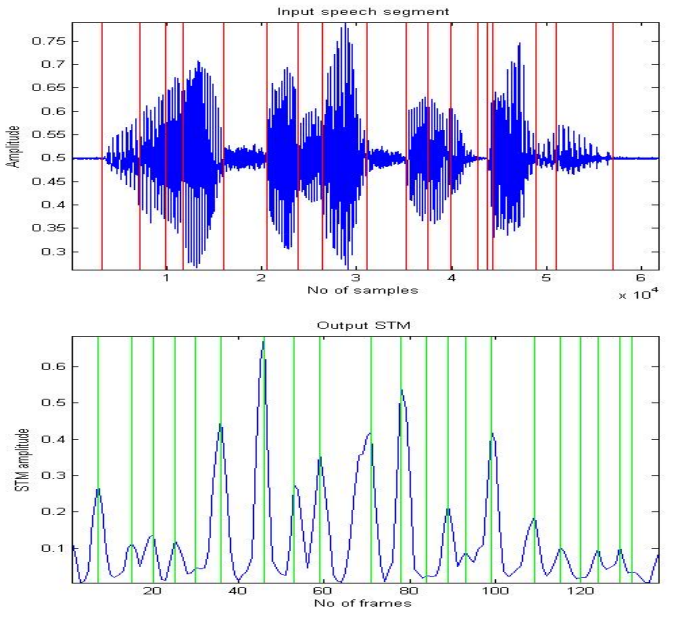


Figure 8.10: (a) manual and (b) automatic segmented phonetic boundary of Marathi sentence, "manushhya naashava.nta aahe".

Table 8.7: # of phonetic boundaries obtained on Gujarati, Marathi and English (TIMIT) database using MFCC_STM

Boundaries Obtained	**Manually**	**Automatic**	**After post processing**
Gujarati	16	19	17
Marathi	19	21	18
English (TIMIT)	15	18	12
TIMIT Train data	219311	310173	221659
TIMIT Test data	85019	121776	87002

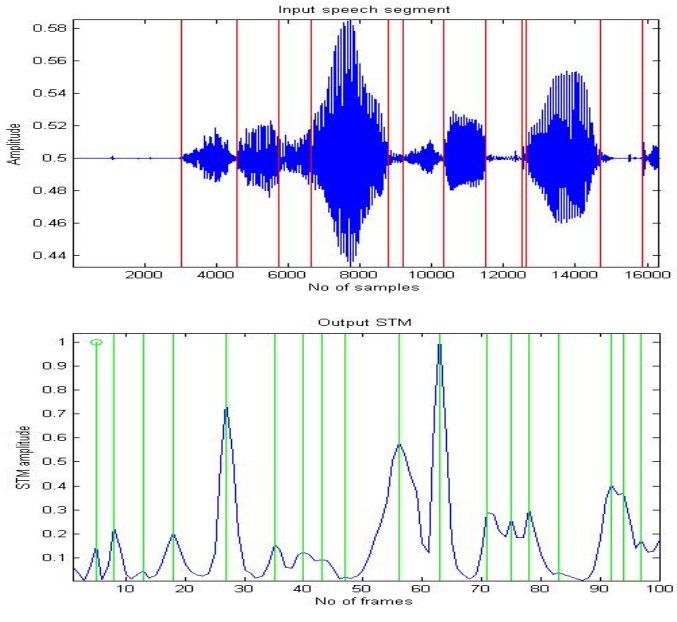


Figure 8.11: (a) manual and (b) automatic segmented phonetic boundary of TIMIT sentence, "she had your dark".

Table 8.8: Performance obtained on entire TIMIT database (in %)

Feature set	***10 ms*** **Agreement**	***15 ms*** **Agreement**	***20 ms*** **Agreement**	***25 ms*** **Agreement**
MFCC_STM [17]	50	68	82	91
CFCC_STM [18]	45	63	76	84
F1	49	68	81	89
F2	52	73	90	100

8.4.2 Automatic broad phonetic classification

First task of this sponsored project is to get manual phonetic transcription of the speech data (with dialect variation) in Gujarati and Marathi languages collected from various dialectal regions of Gujarat and Maharashtra. Our team has collected speech data and started doing phonetic transcription manually as well.

Automatic Phonetic Transcription: Research is going on automatic phonetic transcription of speech signal. Researchers have used various methods to transcribe speech signal automatically; such as some used acoustic-based features, while some used pattern recognition techniques. Some researchers also preferred semi-automatic phonetic transcription, in which they automatically segment speech data at phoneme-level and then manually transcribe it. Generally, the steps followed in this method are as follows: first segment the data at phoneme-level; then find the manner of articulation of the segment; then find the place of articulation of the segment, i.e., a phoneme is identified; and finally, to reduce the errors in automatic transcription, apply HMM or ANN techniques [19]. One of the team member of DA-IICT prosody team, *viz.*, Bhavik Vachhani has completed his M.Tech thesis in the area of *phonetic segmentation*.

Automatic Recognition of Manner of Articulation: Generally researchers prefer temporal features (i.e., noise/frication duration, rise time, rate of rise time, silence duration, etc.) to find the manner of articulation [5]. There has been similar work done in Marathi Language by Vaisali Patil and Preeti Rao from IIT Bombay [20]. It has been found by researchers that there is positive relation between noise/frication duration, rise time and silence duration whereas rate of rise time is negatively related to the former three features [21], [5]. All researchers mainly tried to distinguish the obstruent on the basis of only one of the feature and at most two features. It has also been observed that results varied widely based on the database used as well as the design of database and also position of the obstruent in the word. Comparatively, less work has been done for aspirated and voiced phonemes. One of the team member of DA-IICT prosody team, Kewal D. Malde has completed his M.Tech thesis in the area of *obstruent detection in speech*.

Modulation spectrogram and STM features: Modulation spectrogram-based feature along with simple pattern classifier gives good classification accuracy of obstruents both at broad (i.e., manner or place of articulation) as well as narrow (i.e., phoneme-level). In addition, since for a phoneme, modulation spectrogram looks alike, we can infer that it is speaker-independent representation. Again,

STM-based feature are also explored for the detection of obstruents. Since obstruents are highly dynamic and abrupt in nature, STM contour shoots up while speech transits from obstruent-to-sonorant and sonorant-to-obstruent. This motivates us to use STM information for obstruent detection.

8.4.3 Keyword search in spoken database

Keyword searching in spoken database is an important research problem for audio-based information retrieval [22]. National Institute of Standard and Technology (NIST) has started evaluation for such technology, which is called as Spoken Term Detection STD in 2006 [22]. This can be solved by indexing and searching formulation. During indexing phase, speech database is converted into indices, so as to search the query within indexed version of speech signal. During searching, query is searched within indices and corresponding matching score returns. Some technologies have also built upon the spoken form of query. Audio data and audio query are converted into phonetic units. The matching can be handled by celebrated "Dynamic Time Warping" algorithm [23]. Sliding window-like DTW is used to search the location of audio query. Such techniques are heavily relied onto feature extraction. Hence, feature vectors have to be an excellent representation of linguistic information. One of such representation is posteriorgram wherein each feature vector is converted onto posteriorgram probability vector [24]. Each entry of a vector is the posterior probability of that frame at particular instance for particular phonetic class. The evaluation of such system can be represented in terms of precision. We may select top few candidates based on the detection score obtained. Now, based on the selected candidate, we must get higher precision. It means that selected candidates must corresponds to the actual query.

8.4.4 Classification of Fricative Sounds

Recognition of continuous speech involves finding the phonetic identity of a short-section of speech signal and thereby estimating the phoneme sequence. Classification of short section of speech signal into different phoneme classes (e.g., fricatives *vs.* plosives) based on its acoustic characteristics is an interesting and challenging research problem. In this work, we attempt for classification of one particular class of phonemes, *viz.*, unvoiced fricatives. Fricative sounds are very unique class of phonemes in the sense that for fricatives, the sound source occurs at the point of constriction in the vocal tract rather than at the glottis. Two types of fricatives, *viz.*, voiced and unvoiced have

different speech production mechanisms. In case of voiced fricatives, noisy characteristics caused by the constriction in the vocal tract are accompanied by vibrations of vocal folds, thereby imparting some periodicity into the produced sound. However, vocal folds are relaxed and not vibrating during the production of unvoiced fricatives. This result in lack of periodicity thereby making them difficult to classify based on *spectral* characteristics alone. Voiceless fricatives being noisy, dynamic, relatively short and weak (i.e., low energy) make classification even much more difficult especially in the noisy environments. Since production of unvoiced fricatives is governed by source (e.g., frication noise originating from constriction in vocal tract) - filter (oral cavity) model theory [25], they may be separated depending on location of constriction in oral cavity. This constriction at different locations accounts for distinct acoustical characteristics [26].

8.4.5 Vowel Landmark detection in speech

Landmarks define regions in an utterance around which information about the underlying distinctive features may be extracted. A distinctive feature is the most basic unit of phonological structure. These features are different for every phone. The distinctive features at a landmark are identified based on acoustic measurements made in the vicinity of the landmark [27]. The landmarks and associated features are related to the underlying segments (means bundle of distinctive features corresponding to a particular speech sound unit) and a sequence of segments is hypothesized. This sequence is matched to a lexicon whose words are directly defined in terms of features, and word hypotheses are made. A problem arises with segmentation. However, when parts of the waveform do not have sharp boundaries, like those corresponding to diphthongs (also known as gliding vowel) and semivowels (e.g., /j/ ,/w/). Landmarks are the foci, not the boundaries. This problem of delimiting semivowels and diphthongs is avoided altogether by landmark detection. Landmark detection is typically more hierarchical and involves more than one acoustic measure. In addition, landmarks are associated with bundles of distinctive features whereas segmentation is associated with phones. Lower frequency band information captures glottis excitation information. So the transition obtained from lower frequency zone is helpful for *glottis landmark* detection. A vowel landmark, for instance, is located at the maximum of low frequency energy in the vowel and is used to locate the vowel in a speech. One of the team member, *viz.*, Ankur Undhad of DA-IICT prosody team has been involved in the area of *vowel landmark detection.*

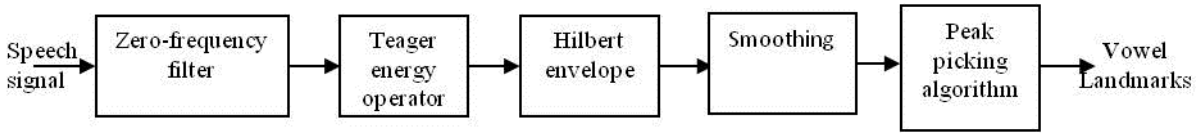


Figure 8.12: Vowel landmark detection algorithm [28].

Fig. 8.12 shows a flow diagram of the algorithm for vowel detection. An input speech signal passes through a *0*-Hz resonator [12], [29]. Then, Teager energy is found of the filtered speech signal. After that Hilbert envelope and smoothing operation are performed on TEO of zero frequency filtered speech signal. Finally, peak picking algorithm is used to detect the nucleus of vowels from that energy profile. Fig. 8.13 shows the syllable nuclie estimated from the algorithm. In addition to the above algorithm,

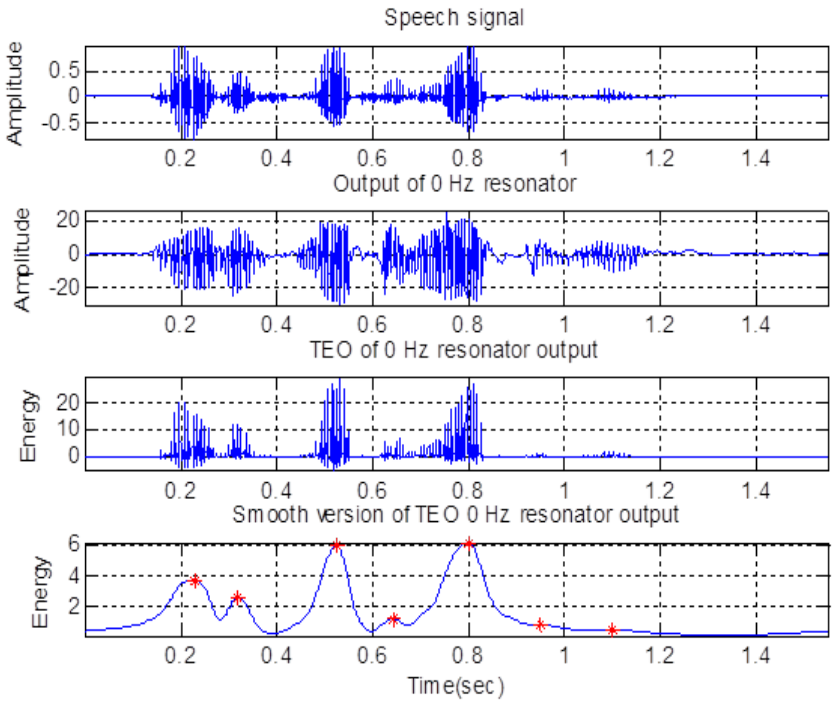


Figure 8.13: An illustration of possible V-landmakrs in speech utterance 'now there's nothing left of me' [28].

two important features are added to quantify vowel landmarks more accurately. Those features are as follows: *1.* zero-crossings rate (ZCR) *2.* minimum vowel duration. For vowel sounds, ZCR is less compared to other sounds. Threshold has to be defined for ZCR and based on that decision will be taken. For a vowel-like sound, minimum duration is around *30 ms*. Output of peak picking algorithm, i.e., peaks corresponds to vowel nucleus are filtered out by these two features. Fig. 8.14 shows that false detection improves.

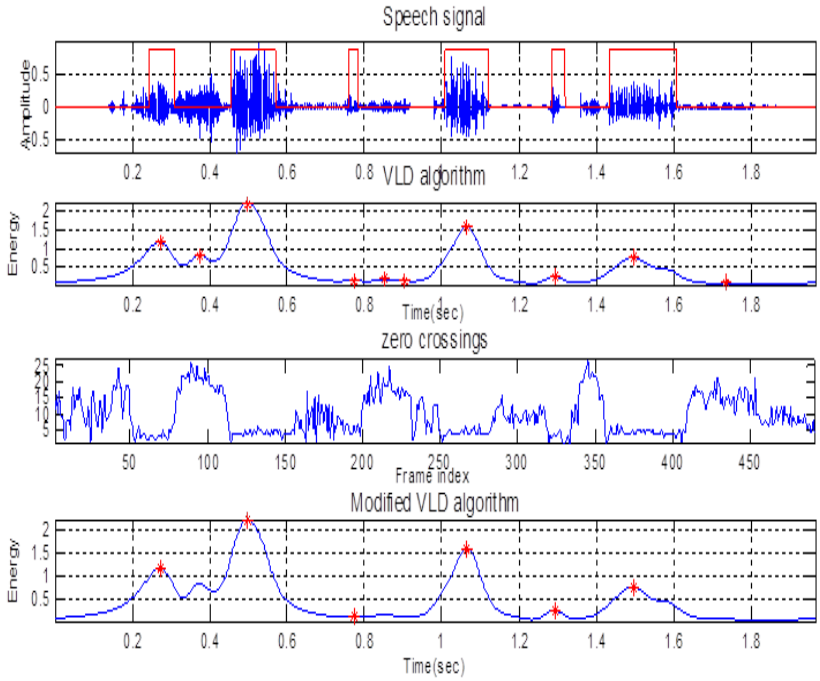


Figure 8.14: An illustration of modified VLD algorithm [28].

8.4.6 Vocal Tract Length Normalization (VTLN):

Even though the variations in acoustic properties in different speakers occur on account of various reasons, it is widely accepted that the major contribution in the variations is by differences in the lengths of vocal tract among speakers. Normalization of vocal tract length (VTL) is a classic problem in the field of speech recognition. A number of methods have been devised for VTLN by various researchers in the past. Andreou *et al.* proposed a maximum-likelihood warp-based method to compensate for the vocal tract length variation and to obtain a set of acoustic vectors that are invariant of the vocal tract length [30]. In this method, the frequency-axis is linearly warped and the implementation is done by resampling of the speech waveform in the time-domain.

Vocal tract length contributes majorly for the variations in acoustic properties in different speakers. The state-of-the-art method for addressing the problem is by modeling the vocal tract as a uniform tube whose formant frequencies vary *inversely* with the vocal tract length with the following relation [31]:

$$F_n = \frac{(2n-1)c}{4L}, \; n \in Z^+, \tag{8.10}$$

where c = velocity of sound, L = length of the tube (i.e., vocal tract). For two speakers, speaker A and speaker B, with different vocal tract lengths L_A and L_B, their respective spectra are scaled versions of each other, i.e., $F_A(\omega) = F_B(\alpha_{AB}\omega)$, where the scaling factor ω_{AB} is the ratio of the vocal tract lengths, i.e., $\alpha_{AB} = L_B/L_A$. Most of the methods make an attempt to estimate appropriate values of α (speaker-specific factor) for all the speakers by maximizing the likelihood [32] or based on the formant frequency [33]. It has been experimentally shown that uniform scaling is not correct and hence, log-warping may not be the appropriate warping function to use to separate the speaker-specific characteristics. It has also been observed that for a given pair of speakers, the scale-factor is different for different vowels and even different for different formants of a vowel. This approach, however, may not be directly useful in speaker-normalization as we have no a priori information about the spoken vowel, formant number and whether the speaker is an adult or a child. Hence, in [31], a corresponding universal frequency-warping function was aimed to be determined, applicable to all the speakers, that helps to separate the speaker-dependencies from the characterization of speech sound. S. Umesh *et al.* proposed the use of scale-cepstral coefficients as features in speech recognition as they provide better separation between the vowels than MFCC [31]. We discussed some methods to normalize vocal tract length by various researchers by finding speaker-specific warping factors. The scale-transform provides a useful tool to achieve speaker normalization without explicitly computing the speaker-specific scaling constant. One of the team member, *viz.*, Shubham Sharma of DA-IICT prosody team was involved in the area of *VTLN*.

8.5 Lab Setup Developed

The various capital equipments, softwares and LDC corpora are purchased as to execute this project work and carry out related research work. The Table 8.9 indicates the equipments, their purpose and quantities.

8.6 Manpower Training

All the consortium meetings were held at IIIT-Hyderabad. Most of the team members of DA-IICT team including principal investigator have attended all the meetings. The Prosody project staff details

Table 8.9: List of equipments purchased, their description, purposes and quantity

Equipment	Description	Purpose	Qty.
Desktop Machines	Dell Optiplex 390 Desktop Computer, Intel Core i5−2400 CPU Processor, 3.10 GHz 3101 MHz FSB 6 MB smart cache 4c, 4 GB (1*4GB)PC3−10600 1333 MHz DDR3 RAM, 1 TB 7200 RPM 3.5" SATA HDD	Computation of algorithm and storage	3
Portable Handy Recorder	Zoom H4n	Field recording	5
Rechargeable cells	Rechargeable cells, Ni-MH AA size, which is used in Zoom H4n	Field recording	20
Dual Earphone with Mic	HP EL283PA	Transcription	9
Matlab with Signal Processing Toolbox	R2012a	Algorithm development	3
LDC Speech Corpora	2002 NIST Speaker Recognition Evaluation (LDC2004S04), TIDIGITS (LDC93S10), SUSAS (LDC99S78), 2008 NIST Speaker Recognition Evaluation Training Set Part 1 (LDC2011S05), 2008 NIST Speaker Recognition Evaluation Test Set (LDC2011S08), NIST Spoken Term Detection (STD) Development Set (LDC 2011S02), NIST Spoken Term Detection (STD) Evaluation Set (LDC 2011S03), 2004 NIST Speaker Recognition Corpus (LDC2006S44), TIMIT (LDC 93S1),NTIMIT (LDC93S2), 2008 NIST Speaker Recognition Evaluation Training Set Part 2 (LDC2011S07), 2008 NIST Speaker Recognition Evaluation Supplemental Set (LDC2011S11)	Computation of algorithm and storage	1
1 TB HDD	External Hard Drive	Storage and data transfer	2
Pen-drive	Transcend 8 GB	Data transfer	11
UPS	APC BR 600CI-IN line Interactive UPS	Power backup	14
Headphone	Bose headphone	speech labelling and listening experiments	2

is as following.

DA-IICT Prosody Staff:

- Maulik Madhavi (Ph.D. Student - Ex Staff member)
- Shubham Sharma (M.Tech Student - Ex Staff member)
- Ankur Undhad (M.Tech - Ex Staff member)
- Bhavik Vacchani (M.Tech - Ex Staff member)
- Kewal Malde (M.Tech - Ex Staff member)

- Tarunima Prabhakar (B.Tech - Ex Staff member)
- Mansi Gokhale (B.Tech - Ex Staff member)
- Vibha Prajapati (B.A. (Eng.), Data Entry Operator)
- Rinni Pandya (Consultants for Transcription)
- Krupa Barot (Consultants for Transcription)
- Bhaveshri Parmar (Consultants for Transcription)
- Maulik Patel (Consultants for Transcription)
- Roma Zala (Consultants for Transcription)
- Gayatri Prajapati (Consultants for Transcription)

Details of Project meeting held are shown in Table 8.10.

8.7 Summary and Future work

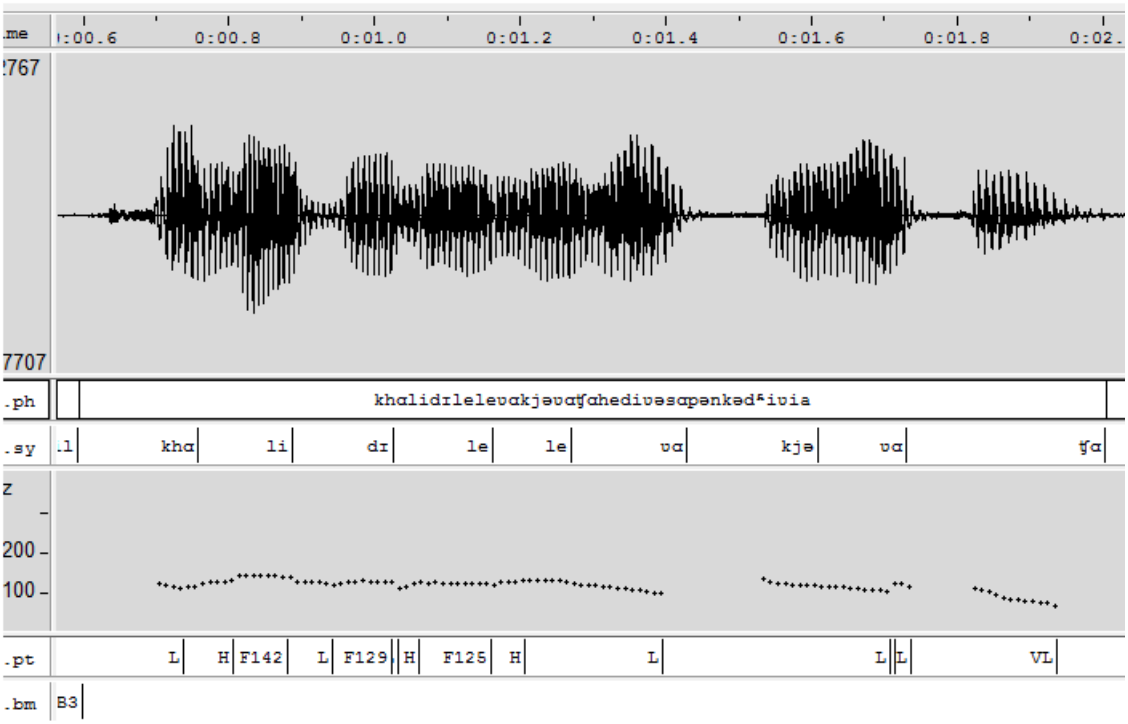
So far DA-IICT team has been involved in the data collection and phonetic transcription of the speech. Based on different recording environment, we observed different behaviour and moods of subjects. It affects the production of speech and hence, its phonetic transcription. In addition, it was observed that while transcribing read speech signal, if speaker speaks appropriately then many times text material information is reflected in phonetic transcription. Our future work may be dedicated towards completion of the data collection from remaining dialectal regions and its phonetic transcription part. Then to find significant events in the speech signal, which may lead to some broad phonetic transcription. The other goal is to design appropriate algorithm for acoustic-phonetic unit segmentation. Finally, raw speech information is converted into phonetic transcript format then one can retrieve the audio-based information from that transcription. This idea may lead to spoken term detection. **Through this DeitY sponsored project, one doctoral student and four M.Tech students are trained and supported for their respective research works and thesis.**

Table 8.10: Manpower Training and Topics Discussed in the Meetings

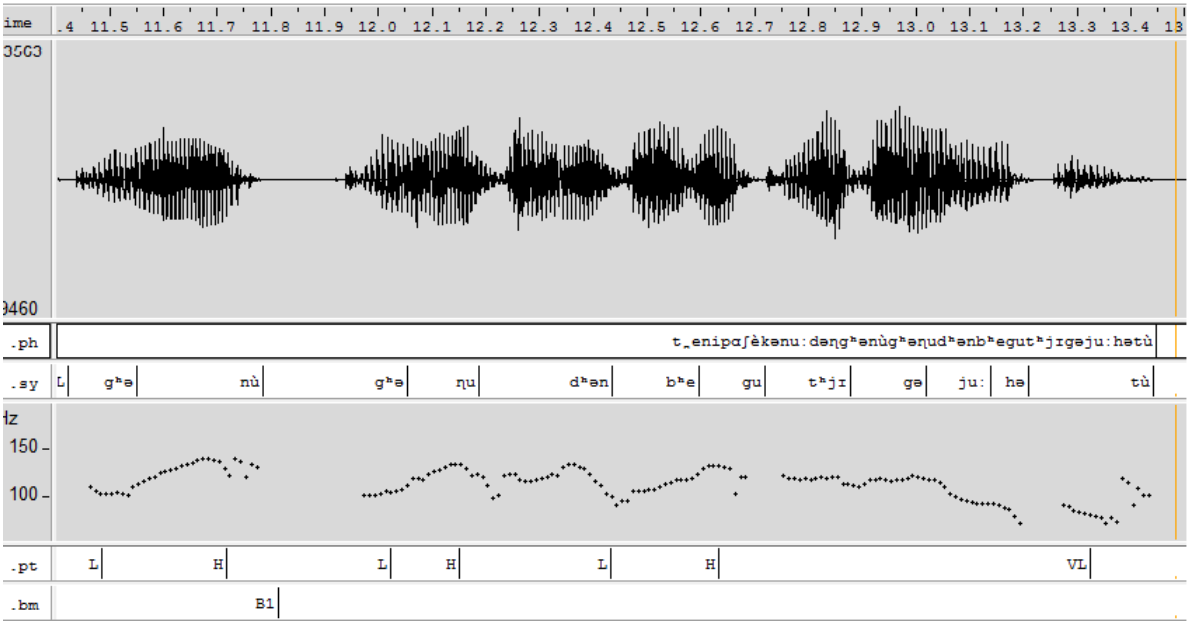
Sr. No	Date of meeting	Participants	Agenda
1	WISP 2011 and Project meeting (Dec, 2011)	Project PI (Prof. Hemant A. Patil)	Project information (IIIT-H)
2	Feb,2012	Prof. Hemant A. Patil, Maulik C. Madhavi, Kewal D. Malde	Use of phonetic transcription in phonetic engine, Concept of phonetic engine. (IIIT-H)
3	May,2012	Prof. Hemant A. Patil, Maulik C. Madhavi, Kewal D. Malde, Bhavik B. Vachhani	Phonetic transcription in detail. (IIIT-H)
4	Oct. 25-28, 2012	Prof. Hemant A. Patil, Maulik C. Madhavi, Kewal D. Malde	Signal level information extraction. Syllabification, pitch marking and break marking. (IIT-H)
5	Dec. 17-18, 2012	Prof. Hemant A. Patil	WISP-2012, Discussion on search engine, Demo of phonetic engine. (IIIT-H)
6	Feb. 5, 2013	Prof. Hemant A. Patil,	PRSG Meeting at DeitY, New Delhi.
7	Mar. 9-10, 2013	Prof. Hemant A. Patil, Maulik C. Madhavi, Kewal D. Malde, Bhavik B. Vachhani, Tarunima Prabhakar, Mansi Gokhale	Speech segmentation, common phoneset and benchmarking data preparation, different techniques for search engine. (Thapar University, Patiala)
8	Oct.12-13, 2013	Prof. Hemant A. Patil, Maulik C. Madhavi, Ankur Undhad, Shubham Sharma	Audio search, demos of phonetic engine, discussion on prosodic marks (pitch and break mark). (DA-IICT, Gandhinagar)
9	Oct.14, 2013	Prof. Hemant A. Patil, Maulik C. Madhavi, Ankur Undhad, Shubham Sharma	One day CEP Workshop on Speech Signal Processing. (DA-IICT, Gandhinagar)
10	Nov. 25-27,2013	Prof. Hemant A. Patil, Maulik C. Madhavi	Oriental COCOSDA 2013 conference. (KIIT Gurgaon)
11	Dec. 10-14, 2013	Prof. Hemant A. Patil	PReMI-2013 conference. (Kolkata)
12	Dec. 13-14, 2013	Prof. Hemant A. Patil	WISP-2013 workshop. (IIT-G)
13	Jan. 17-20, 2014	Prof. Hemant A. Patil, Maulik C. Madhavi, Ankur Undhad, Shubham Sharma	WiSSAP 2014. Deep learning for multilingual speech processing. (IIIT-H)
14	March 7-9, 2014	Prof. Hemant A. Patil, Maulik C. Madhavi	Project review meeting. (IIT-KGP)
15	Sep. 5-6, 2014	Prof. Hemant A. Patil	DeitY project review meeting related to audio search task. (IIT-H).
16	Dec. 12, 2014	Prof. Hemant A. Patil	DeitY project review meeting (IIIT-H)
17	Dec. 13, 2014	Prof. Hemant A. Patil	WISP-2014 workshop. (IIIT-H)
18	Jan. 4-7, 2015	Prof. Hemant A. Patil, Maulik C. Madhavi	WiSSAP 2015. Production-Perception Based New Models of Speech Analysis. (DA-IICT, Gandhinagar)

8.8 Appendix

Marathi Utterance



Guiarati Utterance



Speaker Information

Name (નામ): _____

Age (ઉમર): _____ Gender (જાતિ): _____ Date of Birth (જન્મ તારીખ): _____

Place of Birth (જન્મ સ્થળ): _____

Your upbringing (તમારું રહેઠાણ સ્થળ):_____

Parents Place of Birth (વાલીનું જન્મ સ્થળ): _____

Parent's upbringing (માતા-પિતાનું રહેઠાણ સ્થળ): _____

Native Place (વતન): _____ Mother Tongue (માતૃભાષા):_____

Languages known (જાણીતી ભાષાઓ):_____

Education (અભ્યાસ): _____ Profession (વ્યવસાય): _____

Stay during last one year (છેલ્લા ૧ વર્ષથી ક્યાં રહો છો?): _____

Whether suffered from any of the following diseases (નીચેનામાંથી કોઈ પણ બીમારી હતી/છે)?

- ☐ Surgery of face (મોઢાની સર્જરી)
- ☐ Facial palsy (મોઢાનો લકવો)
- ☐ Face injury (મોઢાની કોઈ ઈજા)
- ☐ Surgery of the throat (ગળામાં સર્જરી)
- ☐ Operation of nasal septum (નાકની સર્જરી)
- ☐ Laryngitis (કંઠનાળનો સોજો)
- ☐ other throat problems (અન્ય ગળાની તકલીફો)

Address (સરનામું): _____

_____

Telephone (ફોન નંબર): _____

E-mail Id: _____

Marital Status (પરણિત/ અપરણિત): _____

If Married, No. of children (If any) (કેટલા બાળકો? જો હોય તો): _____

I am fully aware of the fact that the information given in this form and my speech data are used for the research and academic purpose (Speech Signal Processing). The speech data is collected by project staff members of Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar. This data collection work is a part of sponsored project, *viz*., '*Development of Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages*'. It is supported by Department of Information and Technology (DIT), New Delhi, Government of India. Hence, I don't have any objection regarding this.

હું સંપૂર્ણપણે માહિતગાર છું કે, ફોર્મમાં આપેલ આ જાણકારી અને મારો અવાજ સંશોધન અને શૈક્ષણિક હેતુ (Speech Signal Processing) માટે ઉપયોગમાં લેવામાં આવશે. આ અવાજ, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), ગાંધીનગરના કર્મચારીઓ દ્વારા એકત્ર કરવામાં આવેલ છે. આ એકત્રીકરણ કાર્ય '*Development of Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages'* નામના *sponsored project* ના ભાગરૂપે કરાયેલ છે. ભારત સરકારના Department of Information and Technology (DIT), નવી દિલ્લી હેઠળ હાથ ધરવામાં આવેલ છે. આથી, મને આ સંબંધિત કોઇ સમસ્યા નથી.

Place (સ્થળ): - **Regards** (શુભેચ્છક)

Date (તારીખ):- (Subject's Signature) (સહી)

'*Development of Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages*', sponsored by Department of Information and Technology (DIT), New Delhi, Government of India.

Speaker Information – Marathi Language

Speaker Number: _____, Set Number: _____, Recorder: _____,

File Number: R: _____, I: _____, C: _____, L: _____,

M: _____, W: _____, S: _____, N: _____,

P: _____,

Name (नाव): _____

Age (वय): _____ Gender (लिंग): _____ Date of Birth (जन्म तारीख): _____

Place of Birth (जन्म स्थळ): _____

Your upbringing (पालन पोषणचा स्थळ): _____

Parents Place of Birth (मत पिताचे जन्म स्थळ): _____

Parent's upbringing (मत पिताचे पालन पोषणचा स्थळ): _____

Native Place (वतन): _____ Mother Tongue (मातृभाषा): _____

Languages known (माहिती भाषा): _____

Education (शिक्षण): _____ Occupation (व्यवसाय): _____

Stay during last one year (गेला एक वर्षा तुम्ही कुठे राहतो): _____

Whether suffered from any of the following diseases (या पैकी कोणती बिमारी आहे कां)?

□ Nothing (नाही)

□ Surgery of face (चेहरेचा सर्जरी)

□ Facial palsy (चेहरेचा पक्षाघात)

□ Face injury (चेहरावर इजा)

□ Surgery of the throat (गळाची सर्जरी)

□ Operation of nasal septum (नाकची सर्जरी)

□ Laryngitis (कंठनाळची समस्या)

□ other throat problems (आणखीन काई समस्या)

_____

_____

Address (पत्ता): _____

_____

Telephone (फोन नंबर): _____

E-mail Id: _____

Marital Status (विवाहित): _____

If Married, No. of children (If any) (किती मुल आहेत): _____

I am fully aware of the fact that the information given in this form and my speech data are used for the research and academic purpose (Speech Signal Processing). The speech data is collected by project staff members and Principal Investigator of Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar. This data collection work is a part of sponsored project, *viz*., '*Development of Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages*'. It is supported by Department of Information and Technology (DIT), New Delhi, Government of India. Hence, I don't have any objection regarding this.

मला माहित आहे कि या पत्रकात दिलेली माहिती आणि आवाज हे फक्त संशोधन आणि शैषणिक (Speech Signal Processing) उपयोगासाठी आहे.हि माहिती Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar चा project staff members आणि Principal Investigator करत आहे. '*Development of Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages*' या प्रायोजित प्रकल्पासाठी हि माहिती गोळा कार्नियात येत आहे. Department of Information and Technology (DIT), New Delhi, Government of India, हे या प्रकल्पाचे प्रायोजक आहे. मणून या प्रकल्पाला माझी काही हरकत नाही.

Place (जागा): - **Regards** (आज्ञापालका)

Date (दिनांक):- (Subject's Signature) (सही)

SET-1

વાક્યો

1. હું રોજ સવારે મંદિરમાં જાવ છું.
2. હું ગુજરાતમાં રહું છું.
3. મને ભૂખ લાગી છે.
4. તે આખી રાત સુતો રહ્યો.
5. એ કહેવાની જરૂર નહોતી.
6. આ બગીચામાં જવું નહિ.
7. તમારા ઘરમાં કોણ કોણ છે ?
8. તમે ક્યાં ગામથી આવ્યા છો ?
9. તમે કઈ બાજુ નીકળ્યા હતા ?
10. કેવું સુંદર દ્રશ્ય છે !
11. વાહ! મજા પડી ગયી!
12. કઠોર પરીશ્રમનો કોઈ વિકલ્પ નથી.
13. મુશ્કેલીનો સામનો કરવો તેનું નામ જિંદગી.
14. વિદ્યા પોતે જ એક શક્તિ છે.
15. પવિત્રતા એ જ શાંતિની જનની છે.

શબ્દો

1. ઘઉં
2. બાજરી
3. જુવાર
4. કપાસ
5. મગફળી
6. એરંડા
7. મકાઈ
8. ચોખા
9. જવ
10. જીરું
11. રાઈ
12. તુવેર
13. વટાણા
14. મગ
15. મઠ

આ આકડાંઓને શબ્દોમાં વાંચો .
૦ શૂન્ય
૧ એક
૨ બે
૩ ત્રણ
૪ ચાર
૫ પાંચ
૬ છ
૭ સાત
૮ આઠ
૯ નવ
૧૦ દસ
૧૯- ૩૮-૮૪
૬૬-૮૯-૭૯
૫૫-૬૫-૯૯
૨૮-૬૧-૨૦
૭૨-૪૯-૯૩

તારીખો
૨૮-૧૨-૨૦૧૨
૧૭-૦૫-૧૯૮૫
૯- ૮-૧૯૬૮
૨૦-૨-૧૯૩૯
૧-૦૬-૧૯૯૧

1. કરો કરકસર તો પરિવાર સદ્ધર

કરકસર એ આપણો બીજો ભાઈ છે. એવું હંમેશા ગાંધીજી કહેતા હતાં. જેને આજના મધ્યમ અને ઉચ્ચ મધ્યમ વર્ગે ખૂબ જ સહજતાથી અપનાવ્યું છે. આધુનિક યુગમાં ગાડાના બન્ને પૈંડા એટલે કે પતિ-પત્ની બન્નેએ કમાઈને કુટુંબ નિર્વાહ કરવાનો હોય છે. ઘણી વખત એવું પણ બને છે કે પતિપત્ની બન્ને આપસની સમજૂતીથી અમુક ચોક્કસ સમય સુધી આવકમાંથી કરકસર કરીને બચત કરતાં રહે છે. તેમના દ્વારા નક્કી કરાયેલો સમયગાળો પૂર્ણ થતા, પછી બન્ને વચ્ચે બચત અને કરકસર વિશે વાદ-વિવાદ સર્જાય છે. વર્ષોથી બચતની આદત પડી જતા પતિ હજી પણ મોજશોખ ઉપર કાપ મૂકીને બચતની જ સલાહ આપે છે.

http://www.gujaratsamachar.com/20120703/purti/sahiyar/sahi9.html
Last Update : 03-July-2012, Tuesday

Spontaneous Speech Queries

1. તમારું નામ શું છે ?
2. તમારા ગામનું નામ શું છે?
3. તમારું ગામ ક્યાં જીલ્લામાં આવે ?
4. તમને કઈ ભાષાઓ બોલો.
5. તમારા દિનચર્યા વિશે કહો.
6. તમારા જીવનની યાદગાર પળ વિષે જણાવો.
7. તમારો સ્વભાવ અને જીવન શૈલી વિષે જણાવો.
8. તમારા કામધંધા વિષે જણાવો.
9. તમારા શોખ વિષે કહો.

10. તમારા ગામના લોકો વિષે કહો.
11. ગામના લોકો ક્યાં તહેવારો મનાવે છે ?
12. ગામમાં પાણી, વીજળીની સુવિધા વિષે કહો.
13. બાળકો માટે રમત ના મેદાન અને શાળાઓ
14. ગામના લોકો ખેતી સિવાય બીજા ક્યાં કામધંધામાં સંકળાયેલા છે ?
15. તમારા ગામમાં વાહનની સગવડ ખરા ?
16. તમારા ગામમાં ક્યાંબજાર છે ?

17. તમારા ખેતરમાં ક્યાં પાકો થાય છે ?
18. તમે કયું ખાતર વાપરો છો ?
19. ખેતરમાં કઈ સગવડો અને અગવડો છે ?

20. તમારા ઘરમાં લોકો નો ખોરાક શું છે ?
21. તમારા ઘરમાં કયું તેલ વપરાય છે ?
22. તમને ઘર કામની ચીજ વસ્તુઓ ગામમાંથી મળે છે કે નહિ ?

Read Speech – Marathi Language

Set 1

मराठी माणसाला काय येत ?

मराठी माणसाला भारतीय राज्य घटना लिहिता येते.
मराठी माणसाला पहिला इंडिअन आईदोल बनता येते.
मराठी माणसाला पहिला करोडपती बनता येत.
मराठी माणसाला पहिली नच बलिये विनर बनता येते.
मराठी माणसाला स्वराज्य उभं करता येतं.
मराठी माणसाला भारतीय चित्रपटसृष्टीची मुहूर्तमेढ रोवता येते.
मराठी माणसाला क्रिकेटचा शहेनशहा होता येतं.
मराठी माणसाला महासंगणक बनविता येतो.
मराठी माणसाला पार्श्वगायनात सम्राज्ञी बनता येतं.
मराठी माणसाला संपूर्ण भारतात पहिली मुलींची शाळा काढता येते.
मराठी माणसाला पहिली महिला शिक्षिका बनता येतं.
मराठी माणसाला पहिली महिला डॉक्टर बनता येतं.
मराठी माणसाला पहिली महिला राष्ट्रपती बनता येतं.
** लाभले अम्हांस भाग्य बोलतो मराठी **.

Reference: - http://www.marathimati.com/balmitra/stories/isapniti-katha-1.asp

हे खाली दिलेले शब्द वाचा

सांजवेळ	गरुड	सप्ताह	मिनिट	संध्याकाळ
शाळेतील	शतक	झंझावात	फुप्फुस	हिमवृष्टि
ज्ञानेश्वर	परवा	अंबरठा	महिना	वसंतऋतू
हवामान	स्त्री	चपळ	मांजर	पंधरवडा
कालखंड	म्हैस	जग	पांढरा	रानमांजर
मध्यरात्र	चंडोल	शत्रू	पेंग्विन	सुतारपक्षी

'*Development of Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages*', sponsored by Department of Information and Technology (DIT), New Delhi, Government of India. – Set 1

** खाली दिलेले वाक्य वाचा **

- एवढा अपमान कोण सहन करू शकेल का ?
- कोणीही इतका अपमान सहन करू शकणार नाही.
- त्यांना पार्टीत मजा वाटली नाही का ?
- त्यांना पार्टीत मजा वाटली.
- किती सुंदर दृश्य होतें तें !
- तें दृश्य फार सुंदर होतें.
- मी मंत्री असतों तर !
- मी मंत्री असायला हवा होतों.
- कृपा करून दरवाजा उघडा.
- कृपा करून दरवाजा उघडाल कां ?
- गप बस.
- गप बसशील कां ?
- अमिताभ संजय इतकाच उंच आहे.
- संजय अमिताभ पेक्षां अधिक उंच नाही.
- वर्गात केवल इतका चांगला मुलगा दुसरा नाही.
- केवल वर्गांतला सर्वांत चांगला मुलगा आहे.
- कंचन धीरज इतकी हुशार नाही.
- धीरज कंचनपेक्षां अधिक हुशार आहे.
- धवल आपल्या रोजच्या कामाकडे कधींही दुर्लक्ष करत नाही.
- धवल आपल्या रोजच्या कामाकासे नेहमीं लक्ष देते.

अंकवाचन

० - १ - २ - ३ - ४ - ५ - ६ - ७ - ८ - ९ - १० - १०० - १२५ - १५० - १७५

संख्या
३६ - ६३ - ७८
५९ - २१ - ८४
१८ - ४२ - ९०
८५ - ३४ - ४६
६४ - ८७ - ९९
५६ - ६०० - ७००

क्रमवाचक संख्या
पहिला
दुसरा
तिसरा
चौथा
पाचवा
विसावा

तारीख / दिनांक
१९ - १२ - १९६०
०७ - ०१ - १९६१
०३ - ०७ - १९८९
२६ - ०७ - १९९१
१५ - ११ - २०१३
३१ - १२ - १९४७

अपूर्णांक
पाव
एकचतुर्थांश
अर्धा
एकद्वितीयांश
पाऊण
तीनचतर्थांश

'*Development of Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages*', sponsored by Department of Information and Technology (DIT), New Delhi, Government of India. – Set 1

कहाण्या

चतुराई - गोपाळ भांड

राजा कृष्णचंद्राच्या दरबारांत गोपाळ भांड हा चतुर विदूषक होता. एके दिवशी राजा अगदी उदास झाला होता. गोपाळने त्याबद्दल राजाला विचारल्यावर राजा म्हणाला, "पुन्हां हा नवाबच! "आपल्या गोपाळला आपण सांगणार नाही कां? कदाचित मी आपली मदत करूं शकेन." गोपाळ म्हणाला.

[१]

"नवाबाने मला या टोंकापासून त्या टोंकापर्यंत संबंध पृथ्वीचे माप मोजायला सांगितले आहे. आणि त्याच वेळेला आकाशांतले तारेही मोजायचे आहेत." राजा म्हणाला, "आपल्या गोपाळला कांहीही अशक्य नाही. आपण नवाबाकडे एक लाख रुपये मागावे आणि एक वर्षाची मुदत मागून घ्यावी." विदूषक म्हणाला.

[२]

नवाबासमोर राजा कृष्णचंद्र मोठ्या धीराने उभा राहिला. "तुला एक लाख रुपये हवे आहेत कां? आणि एक वर्षाची मुदत कशासाठीं?" नवाबाने विचारले. "सरकार, मी रात्रंदिवस पृथ्वी मापूं शकतो. पण तारे फक्त रात्रींच दिसतात."

[३]

राजाच्या चेहऱ्यावरचे हसूं गोपाळला दिसले. "हे घे एक लाख रुपये. पण वर्षभरांत काम पूर्ण होईल ना?" राजाने विचारले.

[४]

'*Development of Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages*', sponsored by Department of Information and Technology (DIT), New Delhi, Government of India. – Set 1

[५] [६]

[७] [८]

Reference: http://www.chandamama.com/lang/story/MAR/12/6/0/201/stories.htm

'*Development of Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages*', sponsored by Department of Information and Technology (DIT), New Delhi, Government of India. – Set 1

Semi-Spontaneous Speech – Marathi Language

** प्रश्नोत्तर **

- आपले सुभनाव काय आहे?
- आपल्या आजोबांचे / वडिलांचे नाव काय आहे?
- आपल्या आईचे नाव काय आहे?
- आपले वय किती आहे?
- आपली जन्म तारीख काय आहे?
- आपले जन्म स्थळ कोणते?
- आपले गाव कोणते?
- आपली मातृभाषा कोणती ?
- आपल्याला कोण - कोणती भाषा येते ?
- आपले शिक्षण किती झाले ?
- आपले वाव्साय की आहे ?
- गेल्या एक वर्षापासून आपण कुठे राहते ?
- आपल्या घरी कोण - कोण आहे ?
- हे सगळे काय करतात ?
- आपण दिवसात काय काय कामे करतात ?
- आपल्याला काय आवडते ?
- आपली मनपसंद खाद्यपदार्थ / खेळ / कलाकार ? आणि कां ?
- आपले मार्गदर्शी कोण आहे ? आणि कां ?
- आपले जीवनात आठवण राहिला प्रसंग कोणते ? आणि कां ?
- आपल्या गावं / शहरां विषय माहिती द्या .
- आपल्या गावांत / शहरांत पाणी / विजळी / बँक / ए.ती .एम / अस्पताल / दवाखाना / शाळा / कॉल्लेगे / महाविध्यालेय / मेदान / इंटरनेट / वाहन / बाजार सुविधा कशी आहे?
- आपल्या गावांचे / शहरांचे लोकांचे वाव्साय काय आहेत ?
- आपल्या शेतात कोण कोणती पाक आहे ?
- शेतात कोणते रसायन पयोग होता ?
- शेतात कोण - कोणती सुविधा आहे ?

'*Development of Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages*', sponsored by Department of Information and Technology (DIT), New Delhi, Government of India. QA

Bibliography

[1] A map of Gujarat state,http://upload.wikimedia.org/wikipedia/commons/3/3e/Map_GujDist_Kuchchh.png,(Last accessed on 14 September, 2012).

[2] A map of Maharashtra state,http://www.besttofind.com/img/Map-of-Maharashtra.gif,(Last accessed on 14 September, 2012).

[3] Hemant A. Patil, Maulik C. Madhavi, Kewal D. Malde and Bhavik B. Vachhani, "Phonetic Transcription of Fricatives and Plosives for Gujarati and Marathi Languages," in *Int. Conf. on Asian Language Process., IALP-2012*, pp. 177-180, Hanoi, Vietnam, Nov. 13-15, 2012.

[4] Kewal D. Malde, Bhavik B. Vachhani, Maulik C. Madhavi, Nirav H. Chhayani and Hemant A. Patil, "Development of Speech Corpora in Gujarati and Marathi for Phonetic Transcription," accepted in *16th Int. Oriental COCOSDA conference*, Nov. 25-27,2013.

[5] L. J. Gerstman, "Noise duration as a cue for distinguishing among fricative, affricate, and stop consonants," *The J. of the Acoustical Society of America*, vol. 28, p. 160, 1956.

[6] Furui, S., "On the Role of Spectral Transition for Speech Perception, *J. Acoust. Soc. Amer.*, vol. 80, no. 4, pp. 1016–1025, 1986.

[7] Sorin Dusan and Lawrence Rabiner,"On the Relation between Maximum Spectral Transition Positions and Phone Boundaries," *Proc. INTERSPEECH/ICSLP 2006*, pp. 645–648, 2006.

[8] S. J. Young, et al. "The HTK Book, version 3.4.", Cambridge University Engineering Department, Cambridge, UK, 2006.

[9] T. Nagarajan, H. Murthy and R. Hegde, "Segmentation of speech into syllable-like units," pp. 2893–2896, EUROSPEECH 2003, GENEVA.

[10] Maulik C. Madhavi, Shubham Sharma, and Hemant A. Patil, "Development of language resources for speech application in Gujarati and Marathi," in *Int. Conf. Asian Lang. Process. (IALP)*, pp. 115–118, Kuching, Sarawak, Oct. 20-22, 2014.

[11] A. Sreejith, L. Mary, K. S. Riyas, A. Joseph, A. Augustine, "Automatic prosodic labeling and broad class Phonetic Engine for Malayalam," *Int. Conf. on Control Communication and Computing (ICCC)*, 2013 , pp.522–526, 13-15 Dec. 2013.

[12] B.Yegnanarayana and K. S. R. Murty, "Epoch extraction from speech signals," in *IEEE Trans. Audio, Speech lang. Process*, vol. 16, no. 8, Nov. 2008, pp. 1602–1603.

[13] H. A. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," In *Int. Conf. Acoust., Speech, and Signal Process., (ICASSP'03)*, Vol. 1, pp. I-68-71, 2003.

[14] Y. Zhang, and James R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams, " *IEEE Workshop on. IEEE Automatic Speech Recognition & Understanding, 2009. ASRU 2009*, pp. 398–403, 2009.

[15] P. Ladefoged, "*A Course in Phonetics*", 5^{th}*Ed.*, Boston: Thomson/Wadsworth, 2006.

[16] Venkatesh Keri and Kishore Prahallad, "A comparative study of constrained and unconstrained approaches for segmentation of speech signal,"*INTERSPEECH 2010*, pp. 2238–2241, 2010.

[17] S.B. Davis, and P. Mermelstein,"Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," in *IEEE Trans. on Acoust., Speech, and Signal Process.*, *vol.*28, no. 4, pp. 357-366, 1980.

[18] Qi Li,and Yan Huang,"An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," *IEEE Trans. on Audio, Speech, And Lang. Process.*, vol. 19, no. 6, pp. 1791–181, Aug 2011.

[19] A.M.A. Ali, J. van der Spiegel,P. Mueller,G. Haentjens and J. Berman, "An Acoustic-Phonetic Feature-Based system for Automatic Phoneme Recognition in Continuous Speech," *Proc. of the 1999 IEEE Int. Symposium on Circuits and Systems, ISCAS '99*, vol.3, no., pp.118–121, July 1999.

[20] Vaishali Patil, Preeti Rao, "Acoustic Cues to Manner of Articulation of Obstruents in Marathi," *Proc. of Frontiers of research on Speech and Music (FRSM)*, Kolkata, India, February 2008.

[21] Z. Mahmoodzade and M. Bijankhan, "Acoustic analysis of the Persian fricative-affricate contrast," *Proc. ICPhS XVI*, pp. 921–924, Aug. 2007.

[22] NIST, "The Spoken Term Detection (STD) 2006 Evaluation Plan,", `http://www.itl.nist.gov/iad/mig/tests/std/2006/docs/std06-evalplan-v10.pdf`, 2006.(Last accessed on 14 September, 2012).

[23] H. Sakoe and S. Chiba,"Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. on Acoustics, Speech and Signal Process.*, vol. 26, no. 1, pp.43–49, Feb 1978.

[24] Y. Zhang and J. Glass, "Unsupervised Spoken Keyword Spotting via Segmental DTW on Gaussian Posteriorgrams," *Proc. ASRU*, pp. 398–403, Merano, Dec. 2009.

[25] Stevens, K.N., "Acoustic Phonetics (Current Studies in Linguistics)," *M.I.T. Press*, 1999.

[26] Fant, G, "Acoustic Theory of speech Production," *Mouton*, The Hague, 1960.

[27] S. Liu, "Landmark Detection for Distinctive Feature-Based Speech Recognition," *J. Acoustic Society of America (JASA)*, vol. 100, no. 5, pp. 3417–3430, Nov. 1996.

[28] Ankur Undhad, Hemant A. Patil and Maulik C. Madhavi, "Exploiting speech source information for vowel landmark detection for low resource language," in *9th Int. Symp. Chinese Spoken Lang. Proc., ISCSLP14*, pp. 546–550, Singapore, 12-14 September 2014.

[29] K. Shri Rama Murty, B. Yegnanarayana, and M. Anand Joseph, "Characterization of Glottal Activity From Speech Signals," in *IEEE Signal Processing Letters*, pp. 469–472 , vol. 16, no. 6, June 2009.

[30] A. Andreou, T. Kamm, and J. Cohen, "Experiments in vocal tract length normalization," *Proc. the CAIP Workshop: Frontiers in Speech Recognition II*, 1944.

[31] S. Umesh, L. Cohen, N. Marinovik, and D. Nelson, "Scale transform in speech analysis," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 1, pp. 40–45, January, 1999.

[32] Richard Rose and Li Lee, "A frequency warping approach to speaker normalization," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 1, pp. 40–45, January 1999.

[33] Gish and E. Eide, "'A parametric approach to vocal tract length normalization," *IEEE ICASSP*, vol. 1, pp. 346–349, 1996.

9

IIT Hyderbad

PROGRESS SUMMARY REPORT OF IIT Hyderabad

A. General

A.1	Name of the Project	:	**Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages (Gujarthi & Marathi)**
	Our Reference Letter No	:	11(6)/2011-HCC(TDIL) dated 23-12-2011
A.2	Executing Agency	:	IIT Hyderabad
A.3	Chief Investigator with Designation	:	Dr. K. Sri Rama Murty Assistant Professor
	Co-Chief Investigators with Designation	:	Dr. C. Krishna Mohan Assistant Professor
A.4	Project staffs with Qualification	:	Mr. Phanisankar Nidadavolu (M.Tech from IIT Kanpur, Mr. Naresh Reddy (B.Tech.), Mr. Chandan Behra (B.Tech), Mr. Rafi Mohammad (M.Tech), Mr. Essa Ali Khan (10th Class in Urdu Medium), Mr. Devanuri Prakash (Bachelor of Commerce), Mr. Mettu Srinivas (M.Tech.), Mr. Yaswant Gavani (M.Tech.), Mr. R. S. V. Rao (B.Tech), Ms. Manasa Gadde (B. Tech), Mr. N. Pattabhi Ramaiah (M. Tech), Mr. Malla Sowjya (B. Tech)
A.5	Total Cost of the Project as approved by DIT	:	
	i) Original	:	60.375 Lakhs
	ii) Revised, if any	:	N/A
A.6	Date of starting (Indicate date of first sanction)	:	23 December 2011

A.7	Date of Completion	:	
	i) Original	:	22 December 2012
	ii) Revised, if any	:	31 March 2015
A.8	Date on which last progress report was Submitted	:	5 February 2015

B. Technical

B.1 Works Completed.(detailed report attached) :

- Database collection in three different modes:
  i. Read speech:
  ii. Conversational speech:
  iii. Lecture mode
- Transcribed 10 hours of Telugu data
- Transcribed 10 hours of Urdu data
- Prosodic transcription is done manually
- Developed phonetic engine for Telugu and Urdu
- Developed speech search engine for Telugu and Urdu

B.2 Proposed plan-of-work highlighting the action to be taken to achieve the originally proposed targets :

- Creating GUI for audio search engine
- Evaluating audio search engine on all 12 languages.
- Improving audio search engine using cross-lingual knowledge.

C. Financial

Sl.No	Sanctioned Heads	1st Installment Received	2nd Installment Received	Total Funds Available (A)	Expenditure incurred (23.12.2011 to 31.03.2012) (In Rupees)	Expenditure incurred (01.04.2012 to 31.03.2013) (In Rupees)	Expenditure incurred (01.04.2013 to 31.03.2014) (In Rupees)	Expenditure incurred (01.04.2014 to 31.03.2015) (In Rupees)	Total Expenditure (B)	Balance (In Rupees) A-B
1	Capital Equipment	400000	200000	**600000**	0	400000	0	200000	**600000**	**0**
2	Data Collection	700000	600000	**1300000**	0	619761	0	680239	**1300000**	**0**
3	Consumable stores	200000	100000	**300000**	0	109747	41139	71705	**222591**	**77409**
4	Manpower	1000000	1000000	**2000000**	0	555299	706477	731554	**1993330**	**6670**
5	Travel	100000	100000	**200000**	0	93636	58533	50998	**203167**	**-3167**
6	Workshop &Training	150000	150000	**300000**	0	35814	26500	117576	**179890**	**120110**
7	Contingency	200000	150000	**350000**	0	49322	64113	236372	**349807**	**193**
8	Coordination & Management	100000	100000	**200000**	0	0	0	55707	**55707**	**144293**
9	System Integration	0	0	**0**	0	0	0	0	**0**	
10	Sub Total	**2850000**	**2400000**	**5250000**	**0**	**1863579**	**896762**	**2144151**	**4904492**	**345508**
11	Over Head (15%)	427500	360000	**787500**	0	427500	360000	0	**787500**	**0**
12	**Total Budget**	**3277500**	**2760000**	**6037500**	**0**	**2291079**	**1256762**	**2144151**	**5691992**	**345508**

Principal Investigator

Assistant Registrar (F&A)

Registrar

एन. जयराम
रजिस्ट्रार
N. JAYARAM
REGISTRAR
IIT Hyderabad-502 20

D. Project Outcomes

Publications

1. Pappagari Raghavendra Reddy, and Kallola Rout and K Sri Rama Murty, “Query Word Retrieval From Continuous Speech Using GMM Posteriorgrams", in Proceedings of 2014 International Conference on Signal Processing and Communications (SPCOM), July 2014, Bangalore, India.

2. Pappagari Raghavendra Reddy, and Shekhar Nayak and K Sri Rama Murty, “Unsupervised Spoken Word Retrieval using Gaussian-Bernoulli Restricted Boltzmann Machines", in Proceedings of Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH), September 2014, Singapore

3. Kallola Rout, and Pappagari Raghavendra Reddy and K Sri Rama Murty, “Experimental Studies on Effect of Speaking Mode on Spoken Term Detection", in Proceedings of 2015 National Conference on Communications (NCC), 2015, Mumbai, India.

4. Karthika Vijayan, Vinay Kumar and K. Sri Rama Murty, “Feature extraction from analytic phase of speech signals for speaker verification", in Proceedings of Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH), September 2014, Singapore

5. Karthika Vijayan and K Sri Rama Murty, “Epoch extraction from allpass residual of speech signals", in Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May, 2014, Florence, Italy, pp:1507-1511.

6. Karthika Vijayan and K Sri Rama Murty, “Epoch extraction from allpass residual estimated using orthogonal matching pursuit", in Proceedings of 2014 International Conference on Signal Processing and Communications (SPCOM), July 2014, Bangalore, India.

Databases Developed

- 6 hours of transcribed broadcast news data in Telugu and Urdu
- 2 hours of transcribed conversational data in Telugu and Urdu
- 2 hours of transcribed extempore data in Telugu and Urdu.

Tools & Systems Developed

- Hybrid HMM-ANN based phoneme recognizer was developed for Telugu and urdu
- Spoken-term detection system using phonetic posteriors
- Spoken-term detection system using unsupervised methods, which does not require labelled data.

Appendix 9.1

Detailed Technical Report of IIT Hyderabad

Detailed Technical Report of IIT Hyderabad

0.1 Database collection

IIT Hyderabad is involved in collection and transcription of speech data in Telugu and Urdu languages. Data has been collected from free-to-air broadcast television channels in three different modes: read-speech mode, conversational mode, and extempore mode. The speech data from broadcast news was recorded for read-speech mode. For conversational mode, we recorded data from debates on current affairs involving more than four speakers. Regular lessons taught by primary/high school teachers, in a live class room, are recorded for extempore mode. All the speech data has been recorded at 16kHz sampling frequency from the D2H service installed at Speech Processing lab, IIT Hyderabad.

0.2 Phonetic Transcription

The team at IIT Hyderabad has gained good knowledge on transcription from the workshop conducted at IIIT Hyderabad by Prof. Peri Bhaskara Rao and by Prof. Yegnanarayana. Transcription has been done using the International Phonetic Alphabet (IPA) chart. The data to be transcribed is divided among a group of people. The team listened to the data very carefully and transcribed it. The transcribed data is exchanged among the group for cross validation. This way care has been taken to minimize the errors and to improve the quality of data transcription. A total of 10.5 hours of data is transcribed. Table 5.1 gives the details of the number of hours of data collected, transcribed and verified in each category for both Telugu and Urdu languages. This speech data, along with the manual trascriptions, was used to develop a spoken-term detection system. Rest of this report gives a brief overview of the spoken-term detection system(s) developed at IIT Hyderabad.

Table 2: Amount of data collected and transcribed at IIT Hyderabad

	Telugu			Urdu		
Mocd	Collected	Transcribed	Verified	Collected	Transcribed	Verified
Read	10 h	6 h	6 h	10 h	6 h	6 h
Conversational	5 h	2 h	2 h	5 h	2 h	2 h
Extempore	5 h	2 h	2 h	5 h	2 h	2 h

0.3 Development of Spoken-Term Detection System

It is difficult to manage and monitor increasing volumes of data on the internet. Resources can be efficiently managed, if only the required information can be retrieved from this data. In the case of speech data, we need to search and locate spoken query words in large volumes of continuous speech. This task is termed as Query by Example Spoken Term Detection (STD). Some of the applications of STD include speech data indexing [1], data mining [2], voice dialling and telephone monitoring.

Audio search can be broadly categorized into keyword spotting (KWS) and spoken-term detection (STD), depending on the domain (text or audio) of the query supplied. In KWS system, the query word is supplied as text [3] [4]. Since the query word (text) and reference data (audio) are in two different domains, one of them needs to be transformed into the other domain to perform the search. Hence, the knowledge of the language and pronunciation dictionary is required to implement KWS system. In the case of STD task, the query word is also supplied in the audio format [5], [6]. . While the STD task does not require pronunciation dictionary, it suffers from channel mismatch and speaker variability. One of the main issues in STD is to device a robust and speaker-invariant feature representation for the speech signal, so that the query and reference utterances can be matched in the new representative domain. In this study, we explore methods to obtain robust representation for the STD task.

The STD can be accomplished in two stages: representation and matching. Representation of the speech plays an important role in the STD task, as it is required to differentiate underlying classes. Representation can be obtained through different kinds of data modelling techniques like HMM (Hidden Markov Model), ASM (Acoustic Segment Model), GMM (Gaussian Mixture Model), GBRBM (Gaussian-Bernoulli Restricted Boltzmann Machines), Deep Neural Networks (DNN) etc. In matching stage, test utterance and user query

Table 3: Categorization of STD techniques

	Discrete	Continuous
Supervised	LVCSR [7] Sub-Word modelling [10]	HMM [8], MLP [9], HMM-ANN [5], AKWS [8]
Unsupervised	VQ	GMM posteriors [11], ASM [12], GBRBM [13]

are matched by using different methods depending on the representation of speech. For discrete representation, like in Large Vocabulary Continuous Speech Recognition (LVCSR), lattice matching methods are followed. For continuous representation, template matching methods are employed. An ideal STD system should be able to search for query words quickly without any restriction on the vocabulary and language of query words.

Rest of the reprot is organized as follows: Next section presents a survey of representation and matching approaches used in the state-of-the art STD systems. Advantages and limitations of supervised and unsupervised feature representations is discussed in Section 0.5, . STD systems built using supervised and unsupervised approaches are discussed in Section 0.6 and Section 0.7, respectively. Finally, important contributions of this study are summarized in Section 0.8

0.4 Literature Survey

Most of the methods followed for STD can broadly be categorized into four groups based on the representation. Representation can be in discrete symbols or continuous (template) which in turn obtained by using supervised or unsupervised approaches. Table 3 shows the grouping of the techniques used for speech representation in the context of STD.

0.4.1 Representation of Speech

0.4.1.1 Discrete Representation using Supervised approaches

Large Vocabulary Continuous Speech Recognition (LVCSR) method [14], [7], [15] uses a well trained LVCSR engine to translate the speech into text. Generally, it uses Viterbi search algorithm on word lattice to find the most probable sequence of words based on likelihoods

of acoustic model and language model. It is shown that N-best hypothesis Viterbi search performs better than 1-best hypothesis [7]. Particularly, when Word Error Rate (WER) of the LVCSR system is very high, N-best hypothesis outperforms 1-best hypothesis [16]. Variants of word lattices: Word Confusion Network (WCN) [17], [18], Position Specific Posterior Lattice (PSPL) [19] have been proposed which are more compact and yield better performance for IN-Vocabulary (INV) words. For searching keywords, text based searching methods are employed on translated text. For faster searching speeds, indexing is done on transcribed text [20]. This method is suitable mainly for high resource languages, as it requires large training data specific to language. Since these are word based recognizers, despite the high recognition accuracies for INV, this method suffers from Out-of-Vocabulary (OOV) problem. That is, it can not handle words which are not in the predefined vocabulary. In real-time scenario, encountering OOV word is very common, as many new words are being added everyday to dictionary. So LVCSR based method can not be a practical solution even for high resource languages.

To solve this problem, subword LVCSR based recognizers [10], [21], [14] are proposed, where recognition is done based on subwords instead of words. Since any word can be represented with basic subword set, where subwords can be phonemes, it is possible to recognize any word using subword recognizers. But the performance of subword based recognizers is deteriorated compared to LVCSR based recognizers. To improve the performance, both phonetic and word lattices [16], [22] are considered. Different ways of fusion are considered to combine both lattices. Generally, Sub-word based techniques, words are represented using phonetic lattices. In the matching stage, query search is done by matching lattices.

0.4.1.2 Discrete Representation using Unsupervised approaches

Representing speech using unsupervised approaches in discrete form can be done by using Vector Quantization (VQ). In this method clustering is performed on frames of speech data and a set of mean vectors, called codebook, are stored for representing speech. Feature vectors, derived from speech signal, can be represented as a sequence of codebook indices depending on their proximity to the cluster centers. In the matching stage, the sequence of codebook indices obtained from thre reference and query data can be matched using

approximate substring matching techniques. It's advantages include less complexity, and thereby speed of search engine.

0.4.1.3 Continuous Representation using Supervised Approaches

Acoustic Keyword Spotting (AKWS) method uses two separate models for keywords and non-keywords, namely, keyword model and filler model, respectively. It is based on the principle that likelihood score of keywords calculated from keyword model is higher than the same from filler model, and vice-versa for non-keywords. Then the decision will be made based on their ratio. Ratio of likelihood scores of keyword model and general speech model (background model) can also be considered for the decision. Filler models are generally phoneme loops and keyword models are concatenation of phoneme models. Filler and background models are generally based on Hidden Markov Models (HMM) [8] and neural network [14], [23], [24] techniques which require labelled data. In [25], authors propose ergodic HMM which does not require labelled data. Advantages of a acoustic KWS include simplicity and faster recognition, and hence its feasibility in real-time scenario. Also it does not require any language model. However, the limitation that keywords should be known apriori, limits this method to fixed vocabulary applications, like command controlled devices and telephone routing [23] etc.

In [26], it is shown that posteriorgrams are suitable for STD. Posteriorgram is a template representation of speech segment, unlike text representation. Mathematically, posterior vector P for a speech frame F is defined a

$$P = (P(C_1/F), P(C_2/F), P(C_3/F), ...P(C_M/F)) \tag{1}$$

where $P(C_i/F)$ is probability of frame F belonging to class C_i, and M denotes number of classes. Depending on approach used to extract posteriors, classes can represent phonemes (in HMM), mixtures (in GMM).

In [27], [9], phonetic posteriorgrams are extracted by using well trained phone recognizers. Phonetic posterior vector of a frame is a vector of probabilities belonging to each class. Here each predefined set of phonemes are considered as classes. In [28], phonetic posteriorgrams are extracted using Deep Boltzmann Machines (DBM). It was also shown

that this method is very useful, if limited amount of annotated data is available. In the next stage DTW is applied to search the query template in large speech data archives.

Representing under resourced languages using well built multi-language phone recognizers is also explored by many researchers, and they are briefly summarized in [5], [6]. Features are extracted in two ways: using only the speech data or adapting the supervised phone recognizers built with high resource languages. Phone recognizers are trained with large amount of (labelled) multi-lingual training data so that the trained model can capture underlying phonetic content of speech irrespective of language. Generally these phone recognizers are based on HMM or neural network techniques. A query from any of the low resource languages can also be represented well using trained model, thus overcoming the problem of low resource and identity of language. However, all these approaches require labelled data in at least one language which may not be feasible always.

0.4.1.4 Continuous Representation using Unsupervised Approaches

It is shown in [11], [12] that unsupervised approaches can be potential replacement for phoneme recognizers. In [11], posterior features are obtained from GMM. Modified segmental DTW is applied on query and test utterance posterior features. More informative models, like Acoustic Segment Models (ASM), can be trained by taking account the temporal structure of speech. In this method each phoneme class is represented by HMM, where class labels are obtained from GMM training. An iterative HMM training procedure is followed with the transcriptions obtained in the previous iteration of decoding of HMM. It is shown that ASM outperforms well trained phoneme recognizer in language mismatched environment, and also GMM modelling. In the matching stage, segmental DTW is performed on ASM posteriorgrams. The STD performance has been improved significantly by applying speaker normalization techniques: Constrained Maximum Likelihood Linear Regression (CMLLR), Vocal Tract Length Normalization (VTLN).

0.4.2 Template Matching of Feature Represenations

Statistical behaviour of vocal tract system makes it impossible to generate two exactly similar speech segments in natural speech. So, no two speech utterances have equal dura-

tions of phonemes, and they are distributed non-linearly. In all the two stage approaches, matching stage plays crucial role in bringing out the excerpts of queries from continuous speech. Searching time and memory requirement are two main constraints in matching. Different matching methods are followed based on the representation of the speech. In discrete representation, query words are represented as sequence of symbols or lattices. In [29], three methods for matching lattices are presented: Direct index matching, edit distance alignment and full forward probability. In the case of continuous representation, speech is represented as sequence of frames called as template. Matching templates is attempted in [30] by using DTW. Degree of similarity of two given utterances can be approximated through cost of optimal path of DTW. Limitation of DTW is that durations of two utterances used for matching should be comparable otherwise the computed similarity score denotes the closeness of long utterance with small utterance, like spoken word which are not comparable. Segmental DTW [31], subsequence DTW [32], unbounded DTW [33] and information retrieval DTW [34] are a few of the variants of DTW, which are tuned to search queries in continuous speech. Improvements to IR-DTW by using hierarchical K-means clustering is proposed in [35].

In [31], segmental DTW was proposed for unsupervised pattern discovery. In this technique, starting point in one utterance is fixed, and DTW is applied with a constraint on degree of warping path. Starting point is slid through chosen utterance. In the next step, obtained paths at each point are refined by using length-constrained minimum average (LCMA) algorithm [36] to get minimum distortion segments. In [11], modified segmental DTW, where LCMA is not applied on the paths, is used to accomplish STD goal. As this method requires DTW at periodic intervals of time in either of the two utterances, it requires high computation and memory resources, making it practically impossible on large archives.

In [32], subsequence DTW is proposed for STD task. In this method, calculation of accumulated matrix is different from conventional DTW. In the conventional DTW, dissimilarity values along the reference utterance are accumulated forcing the backtracked path to reach first frame. Using the fact that query can start from any point in reference utterance, accumulation of values along first row of reference utterance is not done, which makes the back-traced path to end at any point in reference utterance. Path is backtracked from minimum accumulated distortion.

In [34], IR-DTW is proposed for the STD task, in which memory requirement was reduced

by avoiding calculation of full distance matrix. Starting indices in both query and reference utterances are chosen, based on exhaustive similarity computation. By using non-linear subsequence matching algorithm, ending indices are found, and the best matching segments are filtered. This algorithm can be scaled up for large amounts of data with reduced memory requirements. When query term contains multiple words, or if query term is not present exactly in the reference utterance, but parts of query term are jumbled in reference utterance (like in QUEEST task in MeidaEval 2015), then this method is useful. This method is further improved by using K-means clustering algorithm instead of exhaustive similarity computation [35]. By imposing constraints on similarity score on matching paths, searching can be made faster as in [37], [38], [39].

Distance measure plays an important role in template matching. Distance between two frames indicates the similarity between them. Small value of distance specifies more similarity. We have experimented with 3 distance measures, namely Euclidean distance, negative logarithm of dot product and Kullback-Leibler divergence. Let T and Q are p^{th} test frame and q^{th} query frame. Then $(p,q)^{th}$ element in distance matrix can be defined using one of the three distance metrics given below:

- Euclidean Distance: It measures the distance between two vectors in Euclidean space.

$$D_{ED}(p,q) = \sqrt{\sum_{k=1}^{M} (T(k) - Q(k))^2}$$

  where M denotes the dimension of the feature vector.

- Negative Logarithm of Dot Product: Geometrically, the dot product of two vectors can be defined as the angle between them, i.e.,

$$D_{DP}(p,q) = \sum_{k=1}^{M} T(k)Q(k).$$

  Dot product is a similarity measure. To use it as a distance measure, negative logarithm of it is taken. So, $(p,q)^{th}$ element can be calculated as

$$D_{LDP}(p,q) = -log(D_{DP}(p,q)).$$

0.4.2.1 Kullback-Leibler Divergence:

Kullback-Leibler Divergence (KL Divergence) is a measure of distance between two probability distributions. In mathematical terms, the KL Divergence from Q to T is defined to be

$$D_{\mathrm{KL}}(T||Q) = \sum_{k=1}^{M} \ln\left(\frac{T(k)}{Q(k)}\right) T(k)$$

This is not a symmetric distance measure. To make it symmetric we took sum of KL divergence between T, Q and Q, T as our distance measure. So, the modified KL divergence is

$$D_{\mathrm{KL}}(p,q) = \sum_{k=1}^{M} \ln\left(\frac{T(k)}{Q(k)}\right) T(k) + \sum_{k=1}^{M} \ln\left(\frac{Q(k)}{T(k)}\right) Q(k)$$

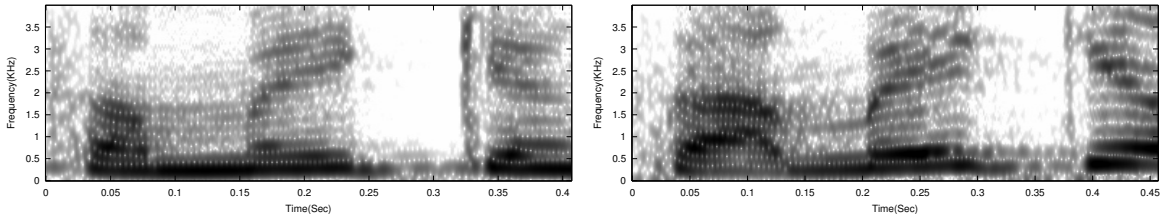
Each vector in posteriorgram can be interpreted as a probability distribution. If any element in posterior vector P is zero then KL divergence with any other vector is infinity, which is not desired. To avoid this, smoothing method is suggested in [9]. So, the new posterior vector P_{new} is

$$P_{new} = (1 - \lambda)P + \lambda\mathcal{U}$$

where $\mathcal{U}$ is uniform distribution, and λ is smoothing constant.

0.5 Representation of Speech for STD

Mel-frequency cepstral coefficients (MFCCs) are the most widely used features in speech recognition systems [40]. Although they are well suited for the statistical pattern matching, like in HMMs for speech recognition, they may not be the best representation for template matching, like in DTW for STD. This behaviour could be attributed to the speaker-specific nature of MFCC features, i.e., they do contain speaker-specific information. Notice that the MFCC features are used for speaker recognition also [41]. Environmental mismatches add to further variations in MFCCs. In the case of statistical pattern matching, the speaker-specific nature of MFCC features is normalized by pooling data from several different speakers. For STD also, we need to derive a stable feature, from the speaker-independent component of



(a) **(b)**

Figure 1: Spectrograms of word "səmaĭkʰjə" spoken by two different speakers

MFCCs, for template matching.

Spectrograms for two instances of word "səmaĭkʰjə" spoken by two different speakers in different contexts are shown in Fig. 1. Formant structure for phoneme /a/ is clearly visible in Fig. 1b, but it is not evident in Fig. 1a. The similarity matrices for same and different speakers across reference utterance and query word by using MFCC are presented in Fig. 3a. The higher the value of an element in the distance matrix, the less similarity between corresponding frames. Elements close to zero are represented with black color. In Fig. 3a at marked area, black path shows more degree of matching in the case of same speaker which is not noticed in the case of a different speaker. For improving performance of STD, there is a need for extracting speaker-independent representation from MFCC features. This can be achieved using statistical models, which capture the underlying speaker-independent distribution of the data.

In this work, we have used posterior representation of speech for developing the STD system. Posterior features are extracted in both supervised and unsupervised approaches. In supervised approach, HMM-ANN hybrid modelling is employed to extract phonetic posteriorgrams using labelled data. In the absence of labelled data, the posterior features are obtained using two unsupervised methods, namely, GMM and GBRBM. Experiments have been conducted on each of these techniques to choose optimal set of parameters. Subsequence DTW is applied on the posterior features to perform query search. Average Precision (P@N) is used as an evaluation metric, which indicates the number of correctly spotted instances P of the word, out of total occurrences N of the word.

0.6 Phonetic Posteriors

Combination of generative and discriminative modelling is proven to be beneficial for classification. Generative models estimate joint density of the input data, while discriminative models capture the boundaries between the classes. Examples for generative models are GMM, HMM, Restricted Boltzmann Machines (RBM) and Gaussian- Bernoulli RBM (GBRBM). Examples for discriminative models include Support Vector Machines (SVM), and Multi Layer Perceptron (MLP). We have used a combination of HMM and MLP for extracting phonetic posteriors.

0.6.1 Speech Segmentation using HMM

HMMs are doubly stochastic models and can be used to model non-stationary signals like speech. Assuming that a state represents a specific articulatory characteristic of speech, the speech signal (observation sequence) can be represented by a sequence of states. Here, state sequence is not known (hidden) to us. Through HMM training, parameters of each state are obtained to capture the statistical properties of speech signal. Generally, each phoneme is modelled using a three-state left-right HMM, assuming that the phoneme is produced in 3 phases. For example, the production of a stop-consonant consist of three main phases, namely, closure phase, burst phase and transition phase into succeeding phoneme. More number of states can also be used for better modelling, but it requires large amount of data for training. In this work, we have used a three-state HMM for each phoneme, where each state is modelled as a 4-mixture GMM. Labelled speech data is used to estimated the parameters of the HMM, i.e., the initial probabilities, emission probabilities and state-transition matrices, using Baum-Welch algorithm [42]. The trained HMM models are used to align the manual transcriptions with the speech data, to obtain the phoneme boundaries. The phoneme boundaries obtained from forced alignment are used for training a Multi Layer Perceptron (MLP).

0.6.2 Extraction of Phonetic Posteriors using MLP

A multilayer perceptron (MLP) is a feedforward artificial neural network model that maps sets of inputs onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, as in Fig. 2, with each layer fully connected to the next one. Each node, in the hidden and output layers, is a neuron (or processing element) with a nonlinear activation function. Sigmoid and hyperbolic tangent activation functions are typically used in the hidden layers, while sofmax activation function is used in the output layer. The output of the softmax function can be interpreted as posterior probability of the class given the input frame. The weights of the network can be learnt, using back-propagation algorithm, by maximizing the cross entropy between the estimated posteriors and actual phoneme labels.

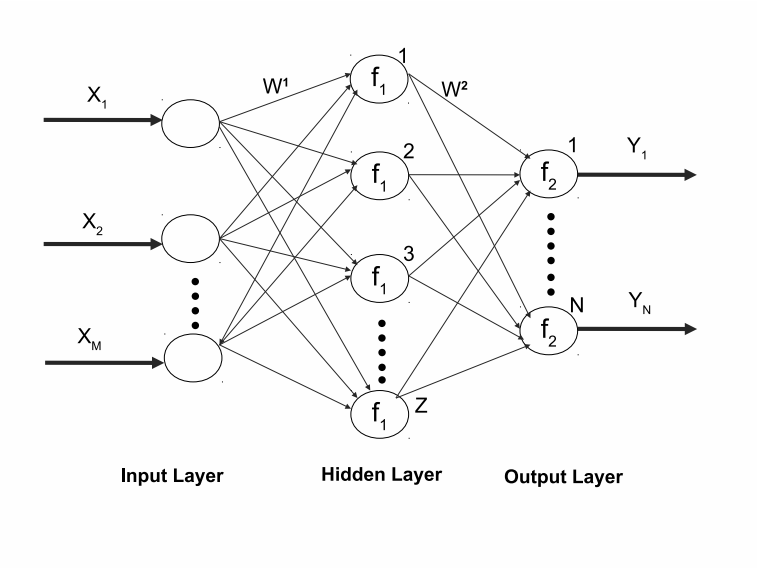


Figure 2: MLP Network, where W^1 is the weight matrix at the input layer and W^2 is the weight matrix at the output layer. X and Y are the input and output vectors respectively. M, Z and N denote the number nodes at the input, hidden and output layer respectively. f_1 and f_2 denotes the activation functions at hidden and output layer respectively.

Unlike GMM, an MLP can be trained with higher dimensional correlated data. Hence, context information can be learnt using MLP by presenting concatenated speech frames as input. In this work, a context of 13-frames was used with 39-dimensional MFCC features to form a 507 $(39 * 13)$ dimensional input feature vector. Phoneme labels obtained from the HMM forced alignment are used as output classes. An MLP with single hidden layer, having 1000 sigmoid units, is trained to map the input feature vectors to the phoneme

Table 4: Recognition accuracy for using different number of phoneme grouping(C)

C	HMM	ANN
6	77.88	81.87
15	70.6	76.02
25	69.51	74.24
45	62.68	69.11

label. The performance of HMM-ANN hybrid approach is evaluated on 5 hours of Telugu broadcast news data, in which 3 hours of data was used for training and the remaining 2 hours was used for testing the system. The performance of HMM-ANN hybrid approach, for different configurations of phonme groupings listed in Table 5, is shown in Table 4. The performance of HMM system alone is also given for comparison. The HMM-ANN hybrid system is consistently better than the HMM alone. As the number of classes increase, there is a gradual drop in the recognition accuracy.

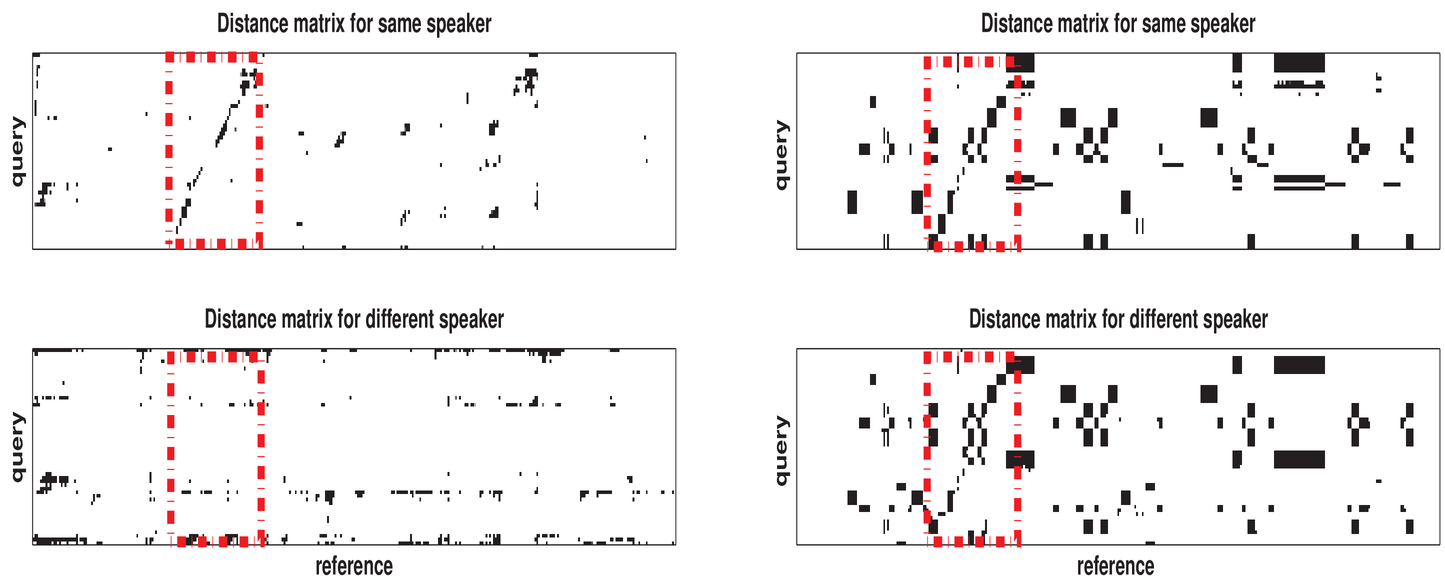
0.6.3 STD using Phonetic Posteriors

The MFCC features extracted from every 25 ms frame of speech signal is converted into phonetic posterior representation using the trained MLP. Since phonetic posteriors are obtained from the MLP, trained with large amount of speech data collected from several speakers, they are more robust to speaker variability. Hence they are better suited for the STD, than the raw MFCC features. Speaker invariant nature of phonetic posteriors is illustrated, in Fig. 3a, using the task of searching for a query word "samaikhya" in the reference utterance "rastram samaikhyamgane unchalantu seemandhranetalu." We have considered two cases: the query word is from same speaker as the reference utterance and the query word is from a different speaker. The distance matrix computed from MFCC features, using Euclidean distance, of reference and query utterances is shown in Fig. 3a(a). Similar matrix computed from posterior features, using KL divergence, are shown in Fig. 3a(b). In the case of matched speakers, there is an unambiguous diagonal path, indicating the presence of query word in the reference utterance, in distance matrices computed from both MFCC and posterior features. When the speakers do not match, the diagonal path is not visible in the distance matrix computed from MFCC features. However, the distance matrix com-

puted from posterior features clearly shows an unambiguous diagonal path. This case study depicts the speaker-invariant nature of posterior features, and thereby their effectiveness in STD.

Subsequence DTW is employed to match the posterior features extracted from the reference and query utterances, and detect the possible matching locations of query in the reference utterance. Performance of STD system is evaluated in terms of average precision (*P@N*), where *N* is number of occurrences of the query in the reference utterance. The evaluation metric P@N is calculated as proportion of query words located correctly *P* in top *N* hits from reference utterance. Detected location is chosen as hit if it overlaps more than 50% with reference location.

The proposed method is evaluated on 2 hours of Telugu broadcast news data with 30 query words spliced from continuous speech data. The performance of the STD system obtained with different phoneme classes is given in Table 6. Even though the phoneme recognition accuracy increased with reducing the number of phoneme classes, it did not result in improved STD. There a is significant decrease in P@N from 15 to 6 classes. Since several phonemes are grouped together into a single class, the number of false matches increases and results in poor performance. Best performance was achieved with 25 phoneme classes, in which the aspirated and unaspirated stop constants produced at the same place of articulation are grouped together. We have used 25 phoneme classes for all the further studies in this work.



(a) **(b)**

Figure 3: Illustration of effectiveness of phonetic posteriors over MFCC. Distance matrix computed from (a-top) MFCC features for matched speakers, (a-bottom) MFCC features for mismatched speakers, (b-top) Phonetic posteriors for matched spakers and (b-bottom) Phonetic posteriors for mismatched speakers

Table 5: Grouping of the phonemes into different classes

<table>
<tr><td>45 classes</td><td>a</td><td>aː</td><td>i</td><td>iː</td><td>j</td><td>u</td><td>uː</td><td>e</td><td>eː</td><td>o</td><td>oː</td><td>ʊ</td><td>ʃ</td><td>ʂ</td><td>s</td><td>h</td><td>f</td><td>m</td><td>n</td><td>k</td><td>k^h</td><td>g</td><td>g^h</td><td>c</td><td>c^h</td><td>z</td><td>j^h</td><td>$ȷ^h$</td><td>ʒ</td><td>ɟ</td><td>ʈ</td><td>$ʈ^h$</td><td>ɖ</td><td>$ɖ^h$</td><td>t</td><td>t^h</td><td>d</td><td>d^h</td><td>p</td><td>p^h</td><td>b</td><td>b^h</td><td>r</td><td>l</td><td>sil</td></tr>
<tr><td>25 classes</td><td colspan="2">a</td><td colspan="2">i</td><td>j</td><td colspan="2">u</td><td colspan="2">e</td><td colspan="2">o</td><td>ʊ</td><td colspan="3">s</td><td>h</td><td>f</td><td>m</td><td>n</td><td colspan="2">k</td><td colspan="2">g</td><td colspan="2">c</td><td colspan="5">j</td><td colspan="2">ʈ</td><td colspan="2">ɖ</td><td colspan="2">t</td><td colspan="2">d</td><td colspan="2">p</td><td colspan="2">b</td><td>r</td><td>l</td><td>sil</td></tr>
<tr><td>15 classes</td><td colspan="2">a</td><td colspan="3">i</td><td colspan="2">u</td><td colspan="2">e</td><td colspan="3">o</td><td colspan="5">F(Fricatives)</td><td colspan="2">N(Nasal)</td><td colspan="4">G(Glottal)</td><td colspan="7">P(Palatal)</td><td colspan="4">R(Retroflex)</td><td colspan="4">D(Dental)</td><td colspan="4">B(Bilabial)</td><td>r</td><td>l</td><td>sil</td></tr>
<tr><td>6 classes</td><td colspan="12">V(Vowel)</td><td colspan="5">F(Fricatives)</td><td colspan="2">N(Nasal)</td><td colspan="23">C(Consonants)</td><td colspan="2">T(Trill and Liquid)</td><td>sil(Silence)</td></tr>
</table>

Table 6: Average performance of STD obtained with different phoneme classes

Metric	6 classes	15 classes	25 classes	45 classes	raw MFCCs
P@N	44.05	77.65	80.13	72.36	45.68
P@2N	55.68	86.50	89.13	80.57	54.91
P@3N	59.17	88.10	90.75	82.39	60.81
P@4N	61.19	88.71	91.28	83.10	63.23
P@5N	62.13	88.99	91.61	83.41	64.29

0.6.4 Effect of Speaking Mode

The experiments, reported in the previous section, were conducted on 30 queries spliced from from continuous read speech. In this section, the performance of the STD system is evaluated on the query words recorded from 20 native Telugu speakers in an isolated manner. It is observed that the duration of the query words recorded in isolated manner is almost the double the duration of those spliced from continuous speech. Since the query words are recorded in a different environment, there is a channel mismatch between the reference and query words. Both these factors (duration and channel mismatch) lead to a significant drop in the STD performance. In order to mitigate with the duration mismatch, we have experimented with the warping path constraints, shown in Fig.4. The weights (a, b, c, d, e) can be used to modify the shape of the warping path. A vertical path can be favoured by decreasing d and e, where as a horizontal path can be favoured by decreasing a and b. A diagonal path can be favoured by decreasing c. In the case of isolated queries, whose duration is much longer, a vertical path should be favoured. Best performance with isolated queries was obtained with the weights (2, 3, 2, 1, 1) In order to normalize the channel effects, cepstral mean and variance normalization was performed for every utterance. The average performance of STD, for both spliced and isolated queries, is given in Table 7. The performance of the STD system with isolated queries is almost 25% less than that of with the spliced queries.

0.7 Unsupervised Posterior Feature Extraction

Even though the performance of supervised methods is satisfactory, they require labelled data to train the models, which may not be available always, particularly for low resource

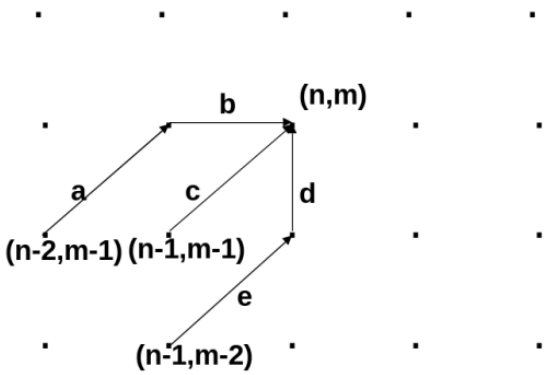


Figure 4: Subsequence DTW path with local weights a b c d e

Table 7: Comparison of STD performance, with queries spliced from continuous speech and queries recorded in isolation

Metric	P@N	P@2N	P@3N	P@4N	P@5N
Spliced from read data	80.49	88.61	90.10	90.67	90.84
Isolated recordings	56.02	66.70	69.66	70.82	71.25

languages. In this scenario, unsupervised approaches can be a promising solution. In this study, two approaches have been presented for posterior feature extraction in the absence of labelled data. Generative models, namely GMM and GBRBM, are employed for unsupervised posterior feature extraction.

0.7.1 Gaussian Mixture Models

Mixture models capture the underlying statistical properties of data. In particular, GMM models the probability distribution of the data as a linear weighted combination of Gaussian densities. That is, given a data set $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, the probability of data X drawn from GMM is

$$p(X) = \sum_{i=1}^{N} w_i \mathcal{N}(X/\mu_\mathbf{i}, \mathbf{\Sigma_i}) \tag{2}$$

where $\mathcal{N}(.)$ is Gaussian distribution, N is number of mixtures, w_i is the weight of the i^{th} Gaussian component, μ_i is its mean vector and $\mathbf{\Sigma}_i$ is its covariance matrix. The parameters of the GMM $\theta_i = \{w_i, \mu_\mathbf{i}, \mathbf{\Sigma_i}\}$ for $i = 1, 2.., N$, can be estimated using Expectation Maximization (EM) algorithm [43]. Given a data point $\mathbf{x}$, the posterior probability that it is generated by

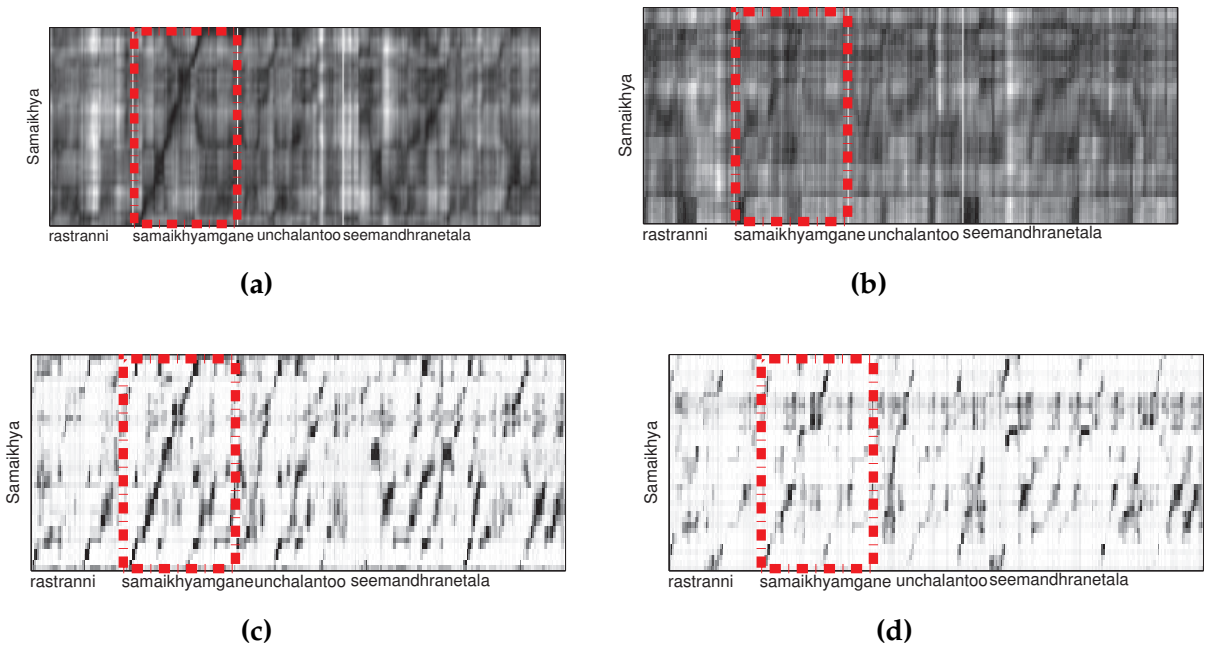


Figure 5: Illustration of effectiveness of GMM posteriorgrams over MFCC. Distance matrix computed from (a) MFCC features for matched speakers, (b) MFCC features for mismatched speakers, (c) GMM Posteriorgrams for matched spakers and (d) GMM posteriorgrams for mismatched speakers

the i^{th} Gaussian component C_i can be computed by using the Bayes' rule as follows:

$$P(C_i/\mathbf{x}) = \frac{w_i \mathcal{N}(\mathbf{x}/\mu_i, \mathbf{\Sigma_i})}{p(\mathbf{x})} \tag{3}$$

The vector of posterior probabilities for $i = 1, 2, \ldots, N$ is called Gaussian posteriorgram.

In this work, we have built a GMM by pooling MFCC feature vectors extracted from 5 hours of speech data. Using this model, the reference and query utterances are represented as a sequence of GMM posteiror features. The distance matrices computed from GMM posteriors, for matched and mismatched speaker conditions, are shown in Fig. 5d(c) and Fig. 5b(d) respectively. The distance matrices computed from MFCC features are also shown in Fig. 5d(a) and Fig. 5b(b). It can be observed from Fig. 5b(b) and Fig. 5b(d) that the GMM posteriorgrams are better at handling the speaker variability.

Subsequence DTW is used to match the GMM posteriorgrams of reference and query utterances, to perform the STD task. The performance of the STD system, with varying number of Gaussian components, is shown in Fig. 6 with KL divergence as distance measure. The

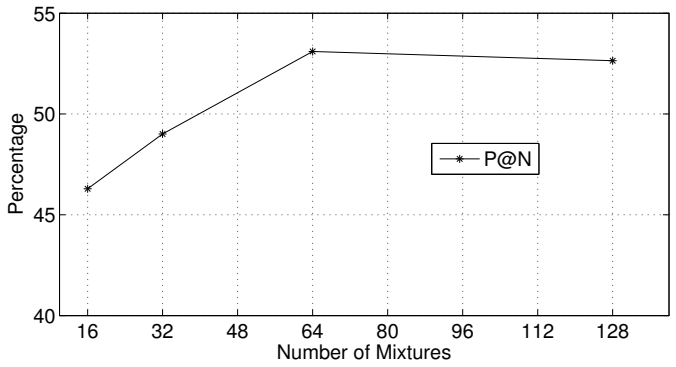


Figure 6: Effect of number of Gaussian mixtures on the performance of STD system

Table 8: Effect of distance measure on the performance of STD system. Performance is evaluated using 64-mixture GMM posteriorgrams

Metric	**MFCC**	**GMM-64**		
	Euclidean Distance	Euclidean Distance	Dot Product	KL Divergence
P@N	45.68%	42.89%	51.89%	53.10%
P@2N	54.91%	50.68%	63.08%	63.69%
P@3N	60.81%	55.67%	67.77%	67.47%
P@4N	63.23%	58.54%	71.55%	70.34%
P@5N	64.29%	60.96%	73.67%	73.67%

performance of the system improved with the number of mixtures. The lower performance of 128-mixture GMM may be attributed to association of each phoneme class with more than one mixture. We adopted 64-mixtures GMM for all the further studies.

The performance of the STD system depends critically on the local distance measure used to compute the distance matrix. In this work, we have experimented with three distance measures, namely, Euclidean distance, negative log cosine distance and KL divergence. The performance of the STD system, with GMM posteriorgrams, for the three distance measures is given in Table 8. While the Euclidean distance is better suited for MFCC features, its performance is not good on GMM posteriors, since the GMM posteriors resemble binary values. The performance of GMM posteriors has improved significantly with negative log cosine distance and KL divergence. In this work, we use KL divergence for the rest of the studies.

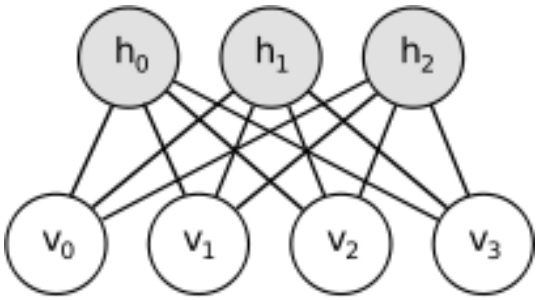


Figure 7: Network architecture of a Restricted Boltzmann Machine

0.7.2 Gaussian-Bernoulli Restricted Boltzmann machine

A Restricted Boltzmann machine (RBM) is an undirected bipartite graphical model with visible and hidden layers. Intra-layer connections do not exit in RBM, as opposed to Boltzmann machine. In an RBM, the output of a visible unit is conditionally Bernoulli given the state of hidden units. Hence the RBM can model only binary valued data. On the other hand in a GBRBM, the output of a visible unit is conditionally Gaussian given the state of hidden units, and hence it can model real valued data. Both in RBM and GBRBM, the output of a hidden unit is conditionally Bernoulli given the state of visible units, and hence can assume only binary hidden states. Since the same binary hidden state is used to sample all the dimensions of the visible layer, GBRBM are capable of modelling correlated data.

A GRBM can be completely characterized by its parameters, i.e., weights, hidden biases, visual biases and variances of the visible units. The GBRBM associates an energy for every configuration of visible and hidden states. The parameters of the GBRBM are estimated such that the overall energy of GBRBM, over the ensemble of training data, reaches a minima on the energy landscape. The energy function for GBRBM, for a particular configuration of real-valued visible state vector $\boldsymbol{v}$ and binary hidden state vector $\boldsymbol{h}$, is defined as

$$E(\boldsymbol{v}, \boldsymbol{h}) = \sum_{i=1}^{V} \frac{(v_i - b_i^v)^2}{2\sigma_i^2} - \sum_{j=1}^{H} b_j^h h_j - \sum_{i=1}^{V} \sum_{j=1}^{H} \frac{v_i}{\sigma_i} h_j w_{ij}, \tag{4}$$

where V and H are total number of visible and hidden units, v_i is the state of i^{th} visible unit, h_j is the state of j^{th} hidden unit, w_{ij} is the weight connecting the i^{th} visible unit to the j^{th} hidden unit, b_i^v is the bias of i^{th} visible unit, b_j^h is the bias of j^{th} hidden unit, σ_i is the variance of the

i^{th} visible unit [44].

The parameters of the GBRBM can be estimated by equating the gradient of the energy function to zero. However, it is not possible to arrive at the closed form solution for the parameters. Contrastive Divergence (CD) algorithm [45] is proposed to estimate the parameters iteratively. In the CD approach, the gradients are estimated as follows

$$\Delta w_{ij} \propto \left(\left\langle \frac{v_i h_j}{\sigma_i} \right\rangle_{data} - \left\langle \frac{v_i h_j}{\sigma_i} \right\rangle_{recall} \right) \tag{5}$$

$$\Delta b_i^v \propto \left(\left\langle \frac{v_i}{\sigma_i^2} \right\rangle_{data} - \left\langle \frac{v_i}{\sigma_i^2} \right\rangle_{recall} \right) \tag{6}$$

$$\Delta b_j^h \propto \left(\left\langle h_j \right\rangle_{data} - \left\langle h_j \right\rangle_{recall} \right) \tag{7}$$

$$\Delta \sigma_i \propto \left(\langle \gamma \rangle_{data} - \langle \gamma \rangle_{recall} \right) \tag{8}$$

where

$$\gamma = \frac{(v_i - b_i^v)^2}{\sigma_i^3} - \sum_{j=1}^{H} \frac{h_j w_{ij} v_i}{\sigma_i^2}$$

and $\langle . \rangle_{data}$ denotes expectation over the input data and $\langle . \rangle_{recall}$ denotes expectation over recalled data.

During each cycle of CD, the energy associated with the joint configuration of visible and hidden states is supposed to decrease, although there is no theoretical guarantee. After a large number of iterations, the expectation of the energy does not change anymore, indicating the thermal equilibrium of the network. At thermal equilibrium, the GBRBM models the joint density of the training data. The trained GRBM model is capable of generating the data points which resemble the training data.

The distribution capturing capability of GBRBM is illustrated, in Fig. 8, with two-dimensional input data. Consider two-dimensional data, marked with blue 'o', in Fig. 8(a). A GBRBM with 6-hidden units is trained, to capture the joint density of this data, for 1000 cycles using CD. The mean of the unbiased samples generated by the trained GBRBM, shown as red'+' in Fig. 8(a), closely follows the original data. The marginal density functions of the

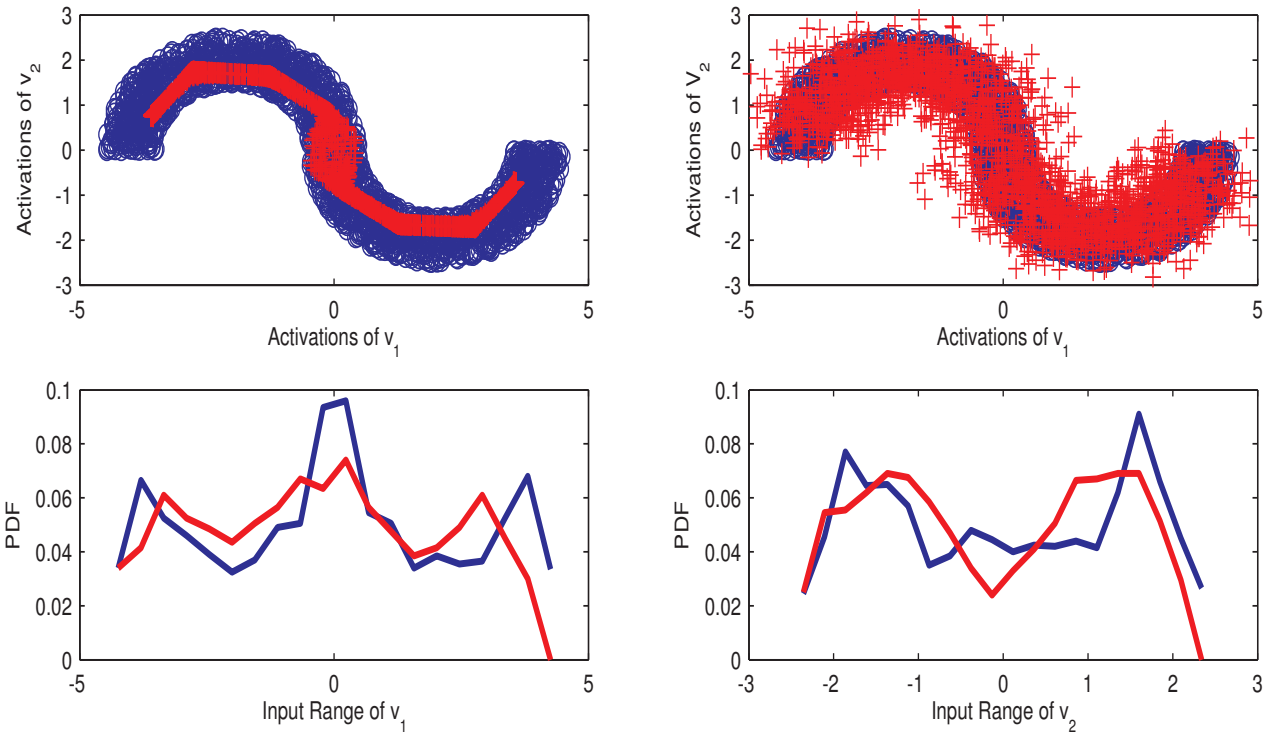


Figure 8: Illustration of distribution capturing capability of GBRBM. (a) Original training data (blue), and mean of the unbiased samples generated by trained GRBM, (b) Original training data (blue), and unbiased samples generated by GRBM, Marginal densities of original (blue) and sampled data (red) along (c) v_1 axis, (d) v_2 axis

original (blue) and estimated (red) data points, along v_1 and v_2 dimensions, are shown in Fig. 8(c) and Fig. 8(d), respectively.

A GBRBM, with 50 hidden units is trained, on 39-dimensional MFCC features to capture the acoustic space spanned by speech data. The marginal densities of the original (blue) and estimated (red) MFCC features is shown in Fig. 9. The marginal densities of the first 12 MFCC features is shown to illustrate the effectiveness of GBRBM in capturing the joint density of the data. In this work, we use the state of the hidden units, at thermal equilibrium, as a feature for STD task. The probability that the j^{th} hidden neuron assumes a state of '1', given the state of the visible units (MFCC features) is computed as

$$P(h_j = 1 \mid \boldsymbol{v}) = sigmoid\left(\sum_{i=1}^{V} \frac{v_i}{\sigma_i} w_{ij} + b_j^h\right) \tag{9}$$

The posterior probability vector, representing the state of all the hidden units, is used to

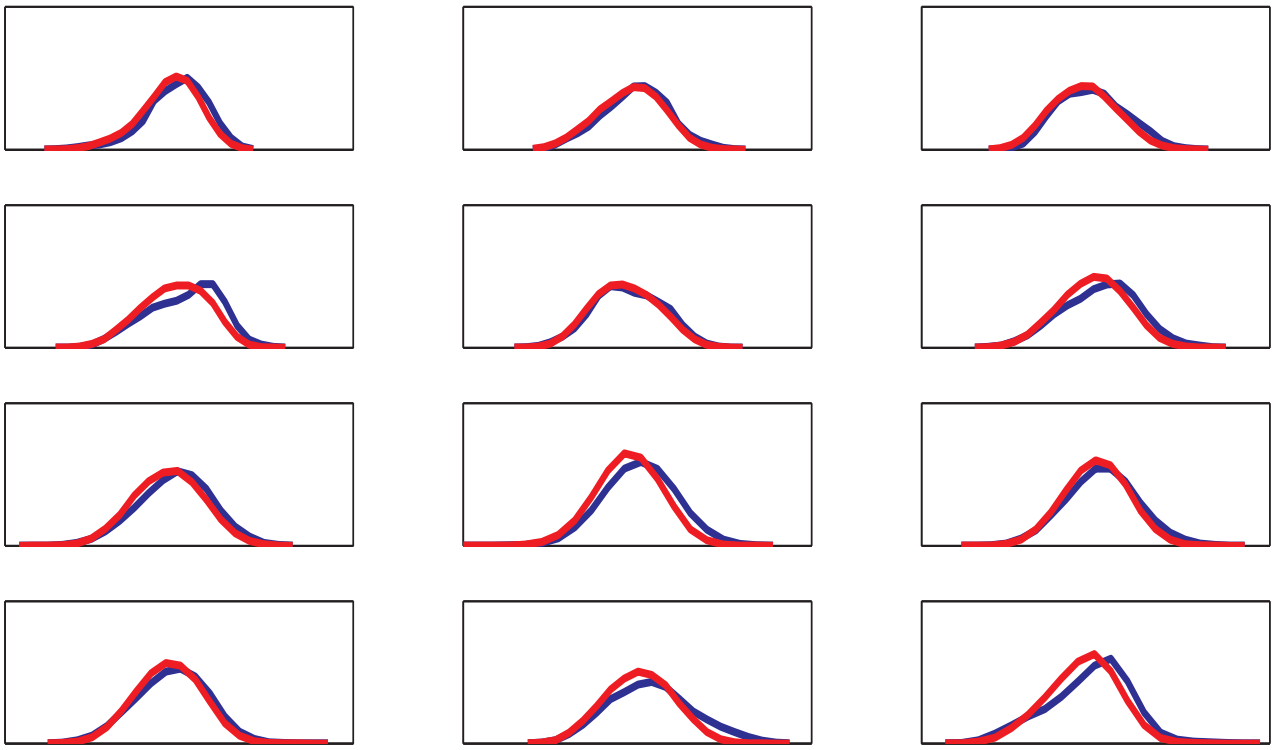


Figure 9: Comparison of original (red) and estimated (blue) marginal densities of first 12 dimensions of MFCC features

match the reference and query utterances.

The distance matrices computed from GBRBM posteriors, with log cosine distance, for matched and mismatched speaker conditions is shown in Fig. 10. In the case of mismatched speakers, the distance matrix computed from GBRBM posteriors clearly shows a diagonal path, in Fig. 10(d), which is absent in distance matrix computed from MFCC features, shown in Fig. 10(b). Hence the GBRBM posteriors is a better representation than MFCC features for STD task. The performance of STD system built with GRBM posteriors is given in Table 9. The performance of the GRBM posteriors is slightly better than the GMM posteriors. Since GMM and GBRBM are two different unsupervised data modeling techniques, the evidences from both these systems are combined linearly. The performance of the combined system is better than the performance of either of the systems alone. However, the performance of the combined system is much lower than the performance of the phonetic posteriors obtained from HMM-ANN hybrid model. This is because the phonetic posteriors are obtained using a supervised approach, while the GMM and GBRBM posteriors are

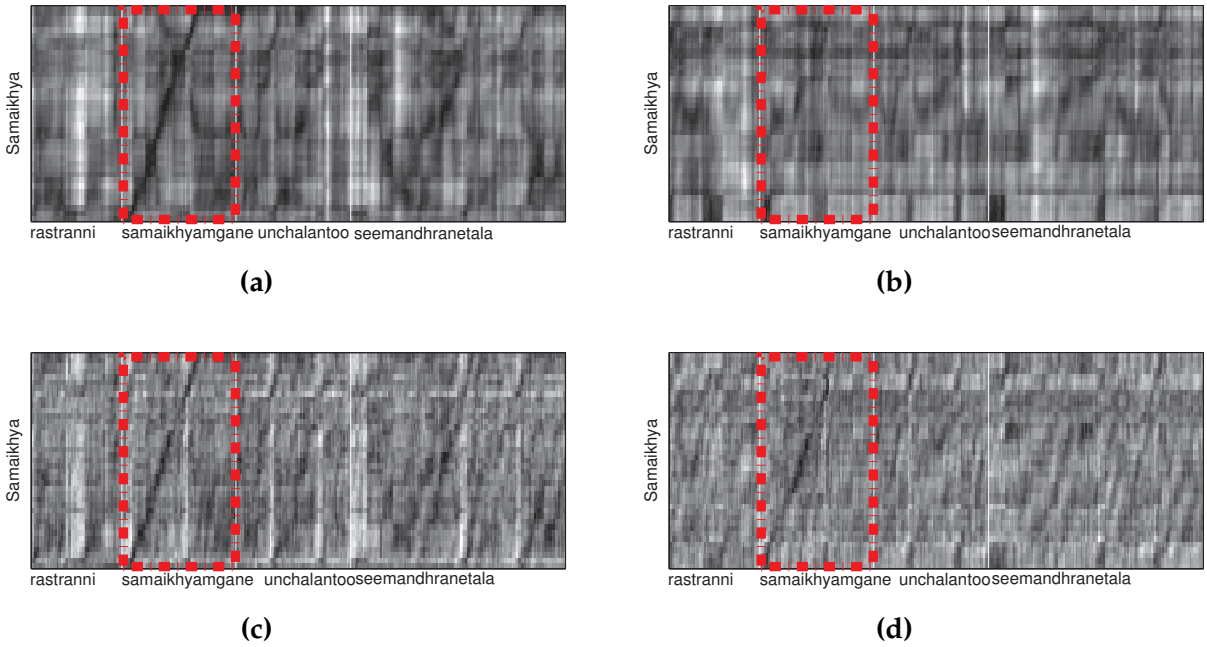


Figure 10: Illustration of effectiveness of GBRBM posteriors over MFCC. Distance matrix computed from (a) MFCC features for matched speakers, (b) MFCC features for mismatched speakers, (c) GBRBM posteriors for matched spakers and (d) GBRBM posteriors for mismatched speakers

Table 9: Performance comparison of STD systems built using different posterior representations

Metric	MFCC	GMM	GBRBM	GBRBM+GMM	HMM-ANN
P@N	45.68%	53.10%	57.64%	59.91%	80.49%

unsupervised approaches. The performance of the STD system, built using different posterior features, on 30 query words from Telugu language is presented in Table 10. Assuming that the probability of misclassification is same for all the syllables, miss rate of longer query words is less compared to smaller query words. On an average, this can be observed in the Table 10 for all representations. For longer query words, the performance is almost similar, with all the three representations, but for smaller query words HMM-ANN posterior features perform better than GMM and GBRBM posteriors.

Table 10: Query words with their P@N(%)

Query word	HMM-ANN	GMM	GBRBM	Query word	HMM-ANN	GMM	GBRBM
prənəbmukʰərji	99.75	100	100	digviɟəjsiŋg	82.78	50	60
teləngaːnə	83.50	90	90	adjəksʰuraːlu	76.96	50	50
səmaːveːsəm	88.97	84.09	86.36	prəbʰutʊəm	71.43	48.57	57.14
saĭləɟaːnaːtʰ	84.64	80	60	adʰikaːrulu	85.31	42.85	64.29
alpəpiːɖənəm	84.62	76.92	69.23	haĭdraːbaːd	83.57	42.85	71.43
paːrləmenʈ	88.24	76.47	76.47	kəmːʈi	58.82	35.29	35.29
bəngaːlaːkʰaːtəm	81.75	75	75	e̥nnikəlu	72.25	35.13	40.54
kaŋgres	78.00	74	78	erpaːʈu	63.38	31.8	40.91
raːɟiːnaːma	85.19	70.37	59.26	vaːtaːvərənəm	67.00	30	50
neːpətʰjəm	62.69	69.23	76.92	vibʰəɟənə	71.43	28.57	33.33
pənchaːjəti	76.62	68.96	82.76	səmaĭkʰjə	93.55	22.58	54.84
soːnijaːgaːndʰi	90.83	66.66	83.33	ɖi̥lliː	50.00	20.83	41.67
poːliŋg	63.75	62.5	75	ʋiʋəraːlu	80.00	20	59.91
kirənkumaːrre̥ɖɖi	95.53	57.14	85.71	ruːpaːji	70.00	20	40
nirnəjəm	83.33	55.55	63.89	məntri	32.73	14.28	12.70

0.8 Summary & Conclusions

In this study, we have presented the development of a spoken term detection system for Indian Languages. Subsequence DTW is employed to search for a query word in the reference utterance. The representation of reference and query utterances plays a crucial role during the search. In this work, we have investigated three different representation techniques, namely phonetic posteriors, GMM posteriors and GBRBM posteriors. The phonetic posteriors, obtained from HMM-ANN phoneme recognizer, requires large amount of manually labelled data. On the other hand, the GMM posteriors and the GBRBM posteriors can be obtained from unlabelled speech data. It was observed that the performance phonetic posteriors is much better than the performance of the GMM and GBRBM posteriors. However, its applications are limited since it requires labelled data. Our future efforts will be focussed on improving the unsupervised feature representation techniques, using sequence and context information.

Bibliography

[1] J. Foote, "An overview of audio information retrieval," *Multimedia Systems*, vol. 7, no. 1, pp. 2–10, 1999.

[2] A. J. Thambiratnam, "Acoustic keyword spotting in speech with applications to data mining," 2005.

[3] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddingtion, "Results of the 2006 spoken term detection evaluation," in *Proc. SIGIR*, vol. 7, 2007, pp. 51–57.

[4] J. Tejedor, D. T. Toledano, X. Anguera, A. Varona, L. F. Hurtado, A. Miguel, and J. ColÃąs, "Query-by-example spoken term detection albayzin 2012 evaluation: overview, systems, results, and discussion." *EURASIP Journal on Audio, Speech and Music Processing*, vol. 23, pp. 1–17, September 2013.

[5] F. Metze, X. Anguera, E. Barnard, M. Davel, and G. Gravier, "The spoken web search task at mediaeval 2012," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8121–8125.

[6] X. Anguera, L. J. Rodriguez-Fuentes, I. Szöke, A. Buzo, F. Metze, and M. Penagarikano, "Query-by-example spoken term detection on multilingual unconstrained speech," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[7] M. Weintraub, "Lvcsr log-likelihood ratio scoring for keyword spotting," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1, May 1995, pp. 297–300 vol.1.

[8] R. Rose and D. Paul, "A hidden markov model based keyword recognition system," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, Apr 1990, pp. 129–132 vol.1.

[9] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 421–426.

[10] K. Ng and V. W. Zue, "Subword-based approaches for spoken document retrieval," *Speech Commun.*, vol. 32, no. 3, pp. 157–186, Oct. 2000.

[11] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 398–403.

[12] H. Wang, C.-C. Leung, T. Lee, B. Ma, and H. Li, "An acoustic segment modeling approach to query-by-example spoken term detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 5157–5160.

[13] R. R. Pappagari, S. Nayak, and K. S. R. Murty, "Unsupervised spoken word retrieval using gaussian-bernoulli restricted boltzmann machines," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014, pp. 1737–1741.

[14] I. SzÃűke, P. Schwarz, P. MatÄŻjka, and M. KarafiÃąt, "Comparison of keyword spotting approaches for informal continuous speech," in *In Proceedings Eurospeech*, 2005.

[15] C. Chelba, T. J. Hazen, and M. SaraÃğlar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Mag*, pp. 39–49, 2008.

[16] M. Saraclar, "Lattice-based search for spoken utterance retrieval," in *In Proceedings of HLT-NAACL 2004*, 2004, pp. 129–136.

[17] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," 2000.

[18] D. Hakkani-Tur and G. Riccardi, "A general algorithm for word graph matrix decomposition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, vol. 1, April 2003, pp. I–596–I–599 vol.1.

[19] C. Chelba and A. Acero, "Position specific posterior lattices for indexing speech," in *In Proceedings of ACL, Ann Arbor*, 2005, pp. 443–450.

[20] D. R. H. Miller, M. Kleber, C.-L. Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, and H. Gish, "Rapid and accurate spoken term detection," in *Annual Conference of the International Speech Communication Association*, 2007, pp. 314–317.

[21] P. Yu, K. Chen, C. Ma, and F. Seide, "Vocabulary-independent indexing of spontaneous speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 635–643, Sept 2005.

[22] D. Can, E. Cooper, A. Sethy, C. White, B. Ramabhadran, and M. Saraclar, "Effect of pronounciations on oov queries in spoken term detection," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 3957–3960.

[23] A. J. K. Thambiratnam, "Acoustic keyword spotting in speech with applications to data mining," Ph.D. dissertation, Queensland University of Technology, 2005.

[24] I. Szöke, M. Fapso, and K. Veselỳ, "But2012 approaches for spoken web search-mediaeval 2012." in *MediaEval*. Citeseer, 2012.

[25] P. Li, J. Liang, and B. Xu, "A novel instance matching based unsupervised keyword spotting system," in *Innovative Computing, Information and Control, 2007. ICICIC '07. Second International Conference on*, Sept 2007, pp. 550–550.

[26] G. Aradilla, J. Vepa, and H. Bourlard, "Using posterior-based features in template matching for speech recognition." in *INTERSPEECH*, 2006.

[27] P. Fousek and H. Hermansky, "Towards asr based on hierarchical posterior-based keyword recognition," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, May 2006, pp. I–I.

[28] Y. Zhang, R. Salakhutdinov, H.-A. Chang, and J. Glass, "Resource configurable spoken query detection using deep boltzmann machines," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 5161–5164.

[29] W. Shen, C. M. White, and T. J. Hazen, "A comparison of query-by-example methods for spoken term detection," DTIC Document, Tech. Rep., 2009.

[30] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, vol. 10, no. 16. Seattle, WA, 1994, pp. 359–370.

[31] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 186–197, 2008.

[32] M. Müller, *Information Retrieval for Music and Motion*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007.

[33] X. Anguera, R. Macrae, and N. Oliver, "Partial sequence matching using an unbounded dynamic time warping algorithm," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, 14-19 March 2010, Sheraton Dallas Hotel, Dallas, Texas, USA*, 2010, pp. 3582–3585.

[34] M. ANGUERA, "Method and system for improved pattern matching," Jun. 25 2014, eP Patent App. EP20,120,382,508. [Online]. Available: http://www.google.com/patents/EP2747078A1?cl=en

[35] G. Mantena and X. Anguera, "Speed improvements to information retrieval-based dynamic time warping using hierarchical k-means clustering," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8515–8519.

[36] Y. ling Lin, T. Jiang, and K. mao Chao, "Efficient algorithms for locating the length-constrained heaviest segments with applications to biomolecular sequence analysis," *Journal of Computer and System Sciences*, vol. 65, pp. 570–586, 2002.

[37] Y. Zhang, K. Adl, and J. Glass, "Fast spoken query detection using lower-bound dynamic time warping on graphical processing units," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 5173–5176.

[38] Y. Zhang and J. R. Glass, "A piecewise aggregate approximation lower-bound estimate for posteriorgram-based dynamic time warping." in *INTERSPEECH*, 2011, pp. 1909–1912.

[39] ——, "An inner-product lower-bound estimate for dynamic time warping," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5660–5663.

[40] N. Alcaraz Meseguer, "Speech analysis for automatic speech recognition," 2009.

[41] M. R. Hasan, M. Jamil, M. Rabbani, and M. Rahman, "Speaker identification using mel frequency cepstral coefficients," *variations*, vol. 1, p. 4, 2004.

[42] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[43] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[44] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Computer Science Department, University of Toronto, Tech. Rep*, 2009.

[45] M. A. Carreira-Perpinan and G. E. Hinton, "On contrastive divergence learning," in *Proceedings of the tenth international workshop on artificial intelligence and statistics*. Citeseer, 2005, pp. 33–40.

10

IIT Kharagpur

Appendix 10.1

Detailed Technical Report of IIT Kharagpur

Development of Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages

Indian Institute of Technology, Kharagpur

10.1 Data Collection

In this work, speech data is collected in three modes: (i) Read speech, (ii) Extempore speech and (iii) Conversational speech in Bengali and Odia languages. Read speech is collected from TV and Radio news bulletins and recordings from a speaker while reading story books and news papers in a closed room with controlled acoustics. Extempore speech is collected from teachers and individuals while speaking on a specific topic. Conversational speech is collected from the casual conversations over mobile phones, live shows in TVs and casual conversations over a table on a particular topic. Zoom H4n portable hand-held voice recorder is used for recording the speech. The speech corpus was recorded using 16 KHz sampling frequency and 16 bits per sample. About 50% of the speech corpus is manually labelled using International Phonetic Alphabet (IPA) symbols. The details of speech corpus are given in Table 10.1.

Table 10.1: Speech corpus details of Bengali and Odia languages.

	Bengali			Odia		
Speech mode	Male (#)	Female (#)	Dur (hrs)	Male (#)	Female (#)	Dur (hrs)
Read	13	28	10	30	30	10
Extempore	7	7	5	9	7	5
Conversation	10	13	5	10	12	5

10.2 Manual Transcription

The speech corpus was Phonetically transcribed manually using IPA symbols. In this work about 64 and 52 IPA symbols are used for transcribing Bengali and Odia speech corpus, respectively. With the help of IPA transcription, one can able to represent basic message as well as prosodic information such as intonation tones, durations and co-articulation effects. In addition to IPA transcription, we

have also carried out syllable boundary markings with manual effort. Transcription of pitch contours and break indices provided by automatic methods is verified manually.

10.3 Automatic Phonetic Transcription

Automatic phonetic transcription was carried out at phone level and syllable level. Phone level phonetic transcription provides the sequence of phones present in the speech utterance. Syllable level transcription provides the sequences of Consonant-Vowel (CV) units present in the speech utterance.

10.3.1 Automatic Phone Level Transcription

Separate Phone Recognition Systems (PRSs) were developed in all three modes of speech for Bengali and Odia languages. The number of phones considered in Bengali for read, extempore and conversation modes of speech are 35, 31 and 31, respectively, while the number of phones considered in Odia are 32, 29 and 28, respectively for read, extempore and conversation modes of speech. Most frequently occurring phones in the IPA transcription are considered for building PRSs. PRSs are developed using Hidden Markov Models (HMMs) and FeedForward Neural Networks (FFNNs). Mel-frequency Cepstral Coefficients (MFCCs) are used as features for building the models. Separate models are developed for Speaker Dependent (SD) and Speaker Independent (SI) cases.

10.3.1.1 Development of Phone Recognition Systems using HMMs

HMM-based systems are developed using a set of context-independent HMMs. A 4-state left-to-right HMM model with a 64 mixture continuous-density diagonal-covariance Gaussian mixture model per state is used to model each sound unit. HMMs are trained using maximum likelihood approach. The global *means* and *variances* are computed from the training data to create flat-start HMMs. The embedded reestimation is carried out on the flat-start HMMs using Baum-Welch algorithm. Viterbi decoding is used for finding the hidden sequence of states within a phone and thereby decoding a speech signal into sequence of phones. The open source HTK toolkit is used for building HMM models.

10.3.1.2 Development of Phone Recognition Systems using FFNNs

We have used three layered FFNNs with linear functionality at the input layer and nonlinear functionality at hidden (second) and output (third) layers. Initially, the frame-level phone labels are

assigned for each speech utterance in the training set. For capturing the hidden relations between MFCC features and the phone labels of the sound unit, the MFCC feature vectors are given as input and information about phone label is given as output during training of the neural network. During training, multiple passes are made through the entire set of training data. Each pass is called an epoch. Initially, we start with a learning rate of 0.008. After each epoch, the performance of the FFNNs is measured with a small set of training data, called the cross validation set, which is held out from main training. The training process will be stopped after the epoch at which the increment in performance improvement is less than 0.5% with cross validation dataset. The advantage of cross-validation based adaptive training scheme is that it provides some protection against over-training. The result of training a FFNN is a set of weights. The softmax nonlinearity activation function is used at output layer to constrain posterior probabilities to lie between zero and one and sum to one. The weights associated to the edges between the nodes can then be used as an acoustic model to convert the features of an unseen test utterance into posterior probabilities of each class. The open source quicknet software is used for training FFNNs. We have used a temporal context of 3 frames with a duration of 45 ms. The number of nodes at input and hidden layer are 117 and 585, respectively.

10.3.1.3 Performance Evaluation of Phone Recognition Systems

The performance of PRSs is determined by comparing the decoded phone labels with the reference transcription of phone labels by performing an optimal string matching using dynamic programming. The number of substitution errors (S), deletion errors (D) and insertion errors (I) are determined for an optimal alignment. Deletion error indicates that, a label is present in the reference transcription but not found in decoded transcription. The substitution error represents that, a label in the reference transcription is substituted with some other label in the decoded transcription. The insertion error indicates that, a label is present in the is decoded transcription but not found in reference transcription. The recognition accuracy in percentage is calculated using Equation 10.1.

$$Percentage\ Accuracy\ =\ \frac{N\text{-}D\text{-}S\text{-}I}{N}\ \times\ 100\ \% \tag{10.1}$$

where N is the total number of labels in the reference transcriptions.

The recognition accuracy of Bengali and Odia PRSs using HMMs and FFNNs are shown in Table 10.2. From the results, it is observed that the Speaker Dependent (SD) systems have better recognition

accuracy compared to Speaker Independent (SI) systems for both the languages. FFNN-based PRSs have higher recognition accuracy compared to HMM-based PRSs. With respect to languages, the phone recognition accuracy of Odia is better than Bengali. For both the languages, the recognition accuracy of read speech is higher than other two modes. Among extempore and conversation modes, the recognition accuracy of extempore speech is better than that of conversation speech.

Table 10.2: Phone recognition accuracy of Bengali and Odia datasets.

Language	Case	Recognition Accuracy (%)					
		HMM			FFNN		
		Read	Extemp	Conver	Read	Extemp	Conver
Bengali	Speaker Dependent	52.14	49.43	33.18	58.85	54.82	39.40
	Speaker Independent	45.48	39.58	37.20	51.20	44.05	32.69
Odia	Speaker Dependent	59.46	55.57	53.30	67.28	62.80	59.88
	Speaker Independent	53.47	47.48	45.80	59.24	56.45	45.81

10.3.2 Automatic Syllable Level Transcription

Automatic transcription at syllable level is carried out using Support Vector Machine (SVM) models. Separate consonant vowel recognition systems (CVRSs) for Bengali and Odia are developed using read speech. MFCC features are extracted using Vowel Onset Point (VOP) as an anchor point. The recognition accuracy of the CVRSs for Bengali and Odia is given in Table 10.3. The performance evaluation is carried out in SD and SI modes. The performance of Odia CVRS is better compared to Bengali CVRS. The recognition accuracy of speaker dependent CV recognition systems have higher recognition accuracy compared to speaker independent systems. The difference between SD and SI modes is about 9% and 28% respectively, for Bengali and Odia. The recognition accuracy for both Bengali and Odia systems is almost same in case of SI mode, whereas for SD mode, the recognition accuracy is very high in case of Odia compared to Bengali.

Table 10.3: CV recognition accuracy of Bengali and Odia datasets.

Case	Bengali	Odia
Speaker Dependent	49.48	69.66
Speaker Independent	40.26	41.59

10.4 Articulatory Features for Phone Recognition

We have explored articulatory features (AFs) for improving the performance of the PRSs developed using read, extempore and conversation modes of speech in Bengali dataset. The AFs are derived from the spectral features using FeedForward Neural Networks (FFNNs). Mel-frequency cepstral coefficients (MFCCs) are used for representing the spectral features. We have considered five AF groups, namely: manner, place, roundness, frontness and height. Five different AF-based tandem PRSs are developed using the combination of MFCCs and AFs derived from FFNNs. Hybrid PRSs are developed by combining the evidences from AF-based tandem PRSs using weighted combination approach.

10.4.1 Extraction of Articulatory Features

The AFs provide crisp representation of each sound unit, in terms of the positioning and movement of various articulators involved in the production of a specific sound unit. AFs varies from one sound unit to another sound unit. Spectral features such as MFCCs capture only the gross shape of the vocal tract, but not the minute variations in the shape of vocal tract. The co-articulation effect between adjacent sound units is captured by AFs. The AFs provide additional clues for discriminating among various sound units. The discrete information about the positioning and movement of articulators with respect to five AF groups is captured. The following subsections describe the details of prediction of AFs using FFNNs.

10.4.1.1 Prediction of Articulatory Features

Table 10.4 shows the articulatory feature specification for read, extempore and conversation modes of speech in Bengali. AF specification represents the possible AF values for each AF group. First column indicates the AF group and the cardinality. The cardinality indicates the number of features in an AF group. Second column lists the possible feature values for each AF group. The AF specification of read speech differs by that of extempore and conversation modes of speech. This is because, the cardinality of *place* AF group of read speech is 9, where as the cardinality of *place* AF group of extempore and conversation speech is 8. Higher cardinality of *place* AF group in read speech is due to the presence of labiodental feature value. The labiodental stands for sounds like /*v*/, but the Bengali speakers have a tendency to use /*bh*/ in place of /*v*/. Hence, the labiodental feature value is not found in *place* AF group of extempore and conversation modes of speech. However, we found very few

instances of labiodental sound units in read speech, which is mainly because of the pronunciations of nouns involving /v/.

Table 10.4: Articulatory Feature Specification for Read, Extempore and Conversation modes of Bengali.

Bengali (Read Speech)	
AF Group (Cardinality)	**Features**
Place (9)	bilabial, labiodental, alveolar, retroflex, palatal, velar, glottal, vowel, silence
Manner (6)	plosive, fricative, approximant, nasal, vowel, silence
Roundness (4)	rounded, unrounded, nil, silence
Frontness (5)	front, mid, back, nil, silence
Height (6)	high, low, mid-high, mid-low, nil, silence
Bengali (Extempore and Conversation modes of Speech)	
AF Group (Cardinality)	**Features**
Place (8)	bilabial, alveolar, retroflex, palatal, velar, glottal, vowel, silence
Manner (6)	plosive, fricative, approximant, nasal, vowel, silence
Roundness (4)	rounded, unrounded, nil, silence
Frontness (5)	front, mid, back, nil, silence
Height (6)	high, low, mid-high, mid-low, nil, silence

The frame-level AFs for each AF group are predicted from the spectral features using AF-predictors. Separate AF-predictors are developed for each AF group. We have explored both HMMs and FFNNs for developing AF-predictors. Figure 10.1 shows the block diagram of prediction of manner AFs. HMM and FFNN-based AF-predictors are developed for manner AF group using MFCCs. The predicted feature values represent the manner AFs.

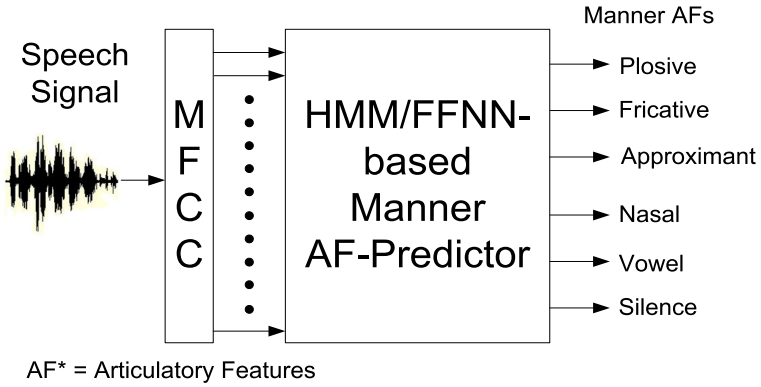


Figure 10.1: Block diagram of Prediction of Manner Articulatory Features.

Similar kinds of AF-predictors are developed for all five AF groups, as shown in Figure 10.2. AFs for a particular AF group are predicted using the AF-predictor of that specific group.

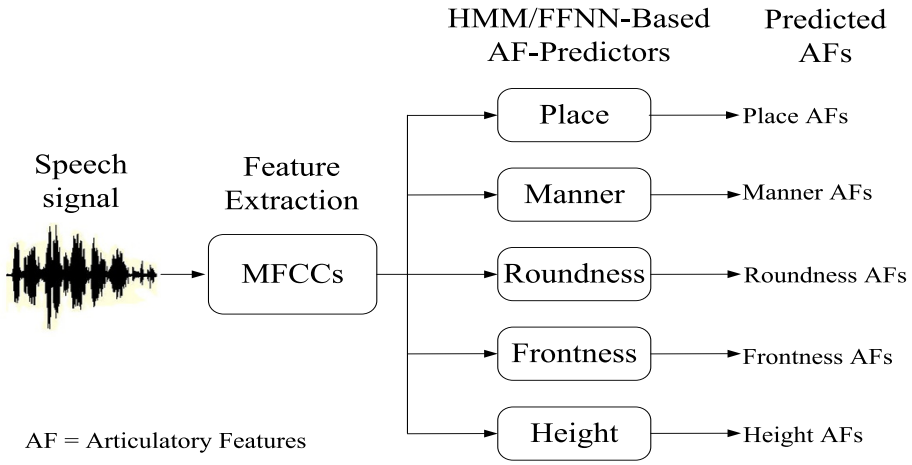


Figure 10.2: Block diagram of the Prediction of Articulatory Features.

For training HMMs and FFNNs, to develop AF-predictors, we require the speech data which is transcribed at AF-level. The AF-level transcription indicates the transcription derived using AF labels. Since the transcription is available at phone level, we derive the AF-level transcription by mapping the phone-labels in the phone-level transcription to AF-labels. An AF label of an AF group represents a possible AF value for that specific AF group. The possible AF labels for each AF group are shown in Table 10.4. The AF-predictors are developed using HMMs and FFNNs using the procedure mentioned in Section 10.3.1.1 and 10.3.1.2, respectively. The size of output layer in developing FFNN-based AF-predictor for a AF group is equal to the cardinality of that AF group as shown in Table 10.4.

Tables 10.5 and 10.6 show the mapping of each phone label into a set of AF labels of various AF groups for read, extempore and conversation modes of Bengali. First column lists unique IPA symbols used in Bengali transcription. Second to sixth columns show the corresponding place, manner, roundness, frontness and height AF values, respectively, for each phone.

10.4.1.2 Performance Evaluation of AF-Predictors

The accuracy of AF-predictors is determined as per the procedure explained in Section 10.3.1.3. Table 10.7 shows the accuracy of prediction of AFs for different AF groups of read, extempore and conversation modes of speech. First column indicates the AF group. Second and third columns show AFs prediction accuracies for read speech, while the fourth and fifth columns tabulates the AFs prediction accuracies for extempore speech. Last two columns show the prediction accuracies for

Table 10.5: Mapping of Phone Labels to AF Groups in for Read Speech in Bengali.

Phones	Articulatory Feature Groups				
	Place	**Manner**	**Roundness**	**Frontness**	**Height**
a	vowel	vowel	unrounded	front	low
o	vowel	vowel	rounded	back	mid-high
ɐ ɜ	vowel	vowel	unrounded	mid	mid-low
i ɪ	vowel	vowel	unrounded	front	high
ɑ	vowel	vowel	unrounded	back	low
ə	vowel	vowel	unrounded	mid	mid-high
ɒ	vowel	vowel	rounded	back	low
u ʊ	vowel	vowel	rounded	back	high
e	vowel	vowel	unrounded	front	mid-high
ɔ	vowel	vowel	rounded	back	mid-low
æ ɛ	vowel	vowel	unrounded	front	mid-low
k k^h g g^h	velar	plosive	nil	nil	nil
ʧ $ʧ^h$ ʤ $ʤ^h$	palatal	plosive	nil	nil	nil
ʈ $ʈ^h$ ɖ $ɖ^h$	retroflex	plosive	nil	nil	nil
t t^h d d^h	alveolar	plosive	nil	nil	nil
p p^h b b^h	bilabial	plosive	nil	nil	nil
m	bilabial	nasal	nil	nil	nil
ɳ	retroflex	nasal	nil	nil	nil
ŋ	velar	nasal	nil	nil	nil
n	alveolar	nasal	nil	nil	nil
s ʃ ʒ	alveolar	fricative	nil	nil	nil
f v	labiodental	fricative	nil	nil	nil
h	glottal	fricative	nil	nil	nil
j	palatal	approximant	nil	nil	nil
ɾ ɹ r l	alveolar	approximant	nil	nil	nil
ɭ	retroflex	approximant	nil	nil	nil
ʋ	labiodental	approximant	nil	nil	nil
sil	silence	silence	silence	silence	silence

Table 10.6: Mapping of Phone labels to AF Groups for Extempore and Conversation modes of Speech in Bengali.

Phones	Articulatory Feature Groups				
	Place	**Manner**	**Roundness**	**Frontness**	**Height**
a	vowel	vowel	unrounded	front	low
ɐ ɜ	vowel	vowel	unrounded	mid	mid-low
ɒ	vowel	vowel	rounded	back	low
ɑ	vowel	vowel	unrounded	back	low
æ ɛ	vowel	vowel	unrounded	front	mid-low
ə ɵ	vowel	vowel	unrounded	mid	mid-high
e	vowel	vowel	unrounded	front	mid-high
œ	vowel	vowel	rounded	front	mid-low
ɞ	vowel	vowel	rounded	mid	mid-low
i ɪ	vowel	vowel	unrounded	front	high
ʏ	vowel	vowel	rounded	front	high
ɔ	vowel	vowel	rounded	back	mid-low
o	vowel	vowel	rounded	back	mid-high
u ʊ	vowel	vowel	rounded	back	high
k k^h g g^h	velar	plosive	nil	nil	nil
ʧ ʧʰ ʤ ʤʰ	palatal	plosive	nil	nil	nil
ʈ $ʈ^h$ ɖ $ɖ^h$	retroflex	plosive	nil	nil	nil
t t^h d d^h	alveolar	plosive	nil	nil	nil
p p^h b b^h	bilabial	plosive	nil	nil	nil
m	bilabial	nasal	nil	nil	nil
ɳ	retroflex	nasal	nil	nil	nil
ŋ	velar	nasal	nil	nil	nil
ɲ	palatal	nasal	nil	nil	nil
n	alveolar	nasal	nil	nil	nil
s ʃ ʒ ɵ ɬ	alveolar	fricative	nil	nil	nil
f v	bilabial	fricative	nil	nil	nil
h	glottal	fricative	nil	nil	nil
x	velar	fricative	nil	nil	nil
ʂ	retroflex	fricative	nil	nil	nil
j	palatal	approximant	nil	nil	nil
ɾ ɹ r l	alveolar	approximant	nil	nil	nil
ɭ	retroflex	approximant	nil	nil	nil
ʋ	bilabial	approximant	nil	nil	nil
sil	silence	silence	silence	silence	silence

conversation speech. The results are shown separately for HMM-based and FFNN-based systems. It is observed that the prediction accuracy of all the AF groups is higher in FFNNs compared to HMMs for read and conversation modes of speech, while the prediction accuracy of most of the AF groups is higher in FFNNs compared HMMs for extempore speech. Since, FFNNs have higher recognition accuracies for all AF groups of read, conversation modes of speech and for majority of AF groups in extempore speech, we have used the FFNNs for predicting the AFs of various AF groups.

Table 10.7: Prediction Accuracy (%) of AF-Predictors of different AF groups across Read, Extempore and Converstaion modes of Speech.

AF Group	**Prediction Accuracy (%) of AF-Predictors**					
	Read		**Extempore**		**Conversation**	
	HMMs	**FFNNs**	**HMMs**	**FFNNs**	**HMMs**	**FFNNs**
Place	55.04	70.35	51.26	62.39	48.72	61.97
Manner	67.51	74.40	63.57	68.19	56.25	65.65
Roundness	68.16	78.58	68.35	65.19	61.58	66.50
Frontness	67.64	74.01	64.37	60.99	58.66	66.48
Height	62.57	67.75	58.30	61.61	55.06	63.17

10.4.2 Prediction of Phone Posteriors

Phone Posteriors (PPs) are predicted from the spectral features using FFNNs. FFNNs perform the phone classification at frame-level. Although HMMs can be used for estimating phone posteriors, FFNNs are employed for this purpose. This is because, FFNNs being discriminative classifiers provide a discriminative way of estimating phone posteriors, while the sequential knowledge capturing ability of HMMs is exploited in later stage of development of PRSs using HMMs. The PPs of phone classes of each frame $p(q_t = i|x_t)$, where q_t is a phone at time t, $i = 1,\ 2\ ...\ N$, and x_t is the acoustic feature vector at time t such that

$$\begin{gathered} \sum_{i=1}^{N} P(i) = 1, \\ where\ N = \ Total\ number\ of\ phone\ classes. \\ i \ = \ indicates\ specific\ phone\ class. \end{gathered} \tag{10.2}$$

FFNN is trained, for predicting the PPs, using the procedure mentioned in Section 10.3.1.2. The weights associated to the edges between the nodes are used as the acoustic model to convert the

features of an unseen test utterance into phone posteriors of each class. Figure 10.3 illustrates the prediction of PPs for ten frames using posteriogram representation. For better visualization of poste-

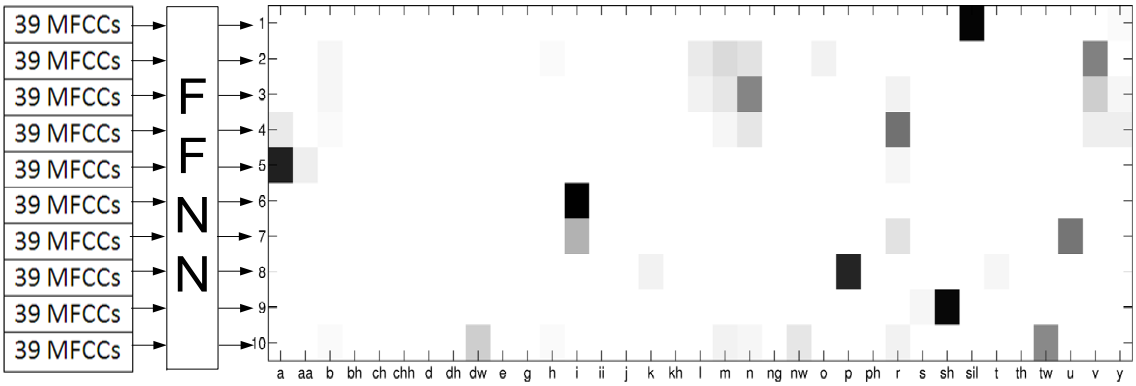


Figure 10.3: Illustration of Prediction of Phone Posteriors for Ten frames using Posteriogram representation.

riogram distribution across all the phones, posteriogram is plotted using non-consecutive frames. The darker spots in the posteriogram indicate higher posterior probability, while the pale spots indicate lower posterior probability. The labels in the *X-axis* of posteriogram indicate the phones used for training the FFNNs. MFCCs extracted from each frame are fed to manner AF-predictor to derive the posteriogram distribution for that specific frame. The sum of all the posterior probabilities obtained for a frame will be equal to 1. The posteriogram distribution represents the PPs. The PPs contain the discriminative knowledge associated to various phonetic units. The dimension of generated PPs will be equal to the number of phones considered for training FFNNs. The number of nodes at input and hidden layer are 117 and 585 units, respectively. The size of output layer is equal to the number of phones considered for training FFNNs.

10.4.3 Development of Tandem Phone Recognition Systems using Articulatory Features

In tandem approach, FFNNs are first trained to perform the classification at frame level, and then the frame-level posterior probability estimates of the FFNNs are used as the acoustic observations in HMMs. The predicted AFs of a particular AF group are augmented with MFCCs to develop AF-based tandem PRS for a specific AF group. Five AF-based tandem PRSs are developed separately, for read, extempore and conversation modes of speech. Table 10.8 shows the phone recognition accuracies of baseline and AF-based tandem PRSs of read, extempore and conversation modes of speech. Figure

10.4 shows the block diagram of manner AF-based tandem PRS. Manner AFs are predicted using manner AF-predictor as shown in Figure 10.1. The predicted manner AFs are combined with MFCCs to develop HMM-based tandem PRS. Similarly, five different tandem PRSs are developed using the predicted AFs from each AF group.

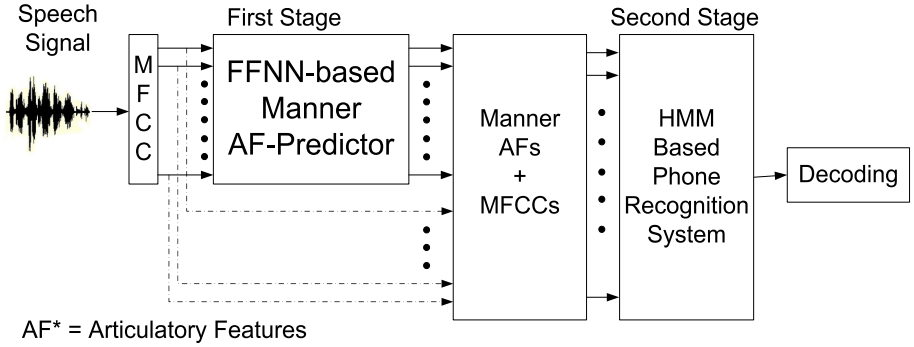


Figure 10.4: Block diagram of the Manner AF-based tandem PRS.

Table 10.8 shows the phone recognition accuracies of baseline and AF-based tandem PRSs of read, extempore and conversation modes of speech. First column shows the different types of features used in development of PRSs. Second, third and fourth columns indicate the recognition accuracies obtained using read, extempore and conversation modes of speech, respectively. It is observed that all AF-based tandem PRSs have higher recognition accuracy compared to baseline PRSs in all three modes of speech. Among vowel AF groups, the *height* AF-based tandem PRSs have shown higher recognition accuracy in all the three modes of speech. Among consonant AF groups, the *place* AF-based tandem PRSs have shown higher recognition accuracy in all the three modes of speech. *Place* AF-based tandem PRSs of read and conversation modes of speech have highest recognition accuracy, whereas the *height* AF-based tandem PRS has highest recognition accuracy in extempore mode of speech.

10.4.4 Hybrid Phone Recognition Systems using Articulatory Features

The hybrid PRSs are developed by combining AF-based tandem PRSs using weighted combination approach. Hybrid systems are developed by using the following combinations of AF-based tandem PRSs : i) place and manner ii) roundness, frontness and height iii) place, manner, roundness, frontness and height (i.e. all AF-based tandem PRSs). As the place and manner AFs mainly capture the characteristics of consonants, the hybrid PRSs developed using place and manner AF-based tandem

Table 10.8: Phone recognition accuracy (%) of AF-based Tandem PRSs across Read, Extempore and Conversation modes of Speech.

Features	Recognition Accuracy (%)		
	Read	Extempore	Conversation
MFCCs (Baseline)	45.48	39.58	37.20
MFCCs + Place AFs	48.89	42.15	40.66
MFCCs + Manner AFs	47.74	41.11	40.18
MFCCs + Roundness AFs	47.28	40.46	38.45
MFCCs + Frontness AFs	46.59	40.75	38.85
MFCCs + Height AFs	48.60	42.93	39.40

PRSs are called Consonant-AF-based hybrid PRSs. Since the roundness, frontness and height AFs mainly capture the characteristics of vowels, the hybrid PRSs developed using roundness, frontness and height AF-based tandem PRSs are called Vowel-AF-based hybrid PRSs. The hybrid PRSs developed using combination of all the five AF-based tandem PRSs are called All-AF-based hybrid PRSs. PP-based tandem PRSs are developed to compare the performance of AF-based hybrid PRSs with PP-based tandem PRSs. The PPs are predicted as per the procedure mentioned in Section 10.4.2. The combination of MFCCs and PPs is used for developing PP-based tandem PRSs using HMMs.

Figure 10.5 shows the block diagram of development of hybrid PRSs. MFCCs are combined with the predicted AFs of each AF group to develop tandem PRSs for each AF group. The scores from all the five tandem PRSs are combined using weighted combination approach.

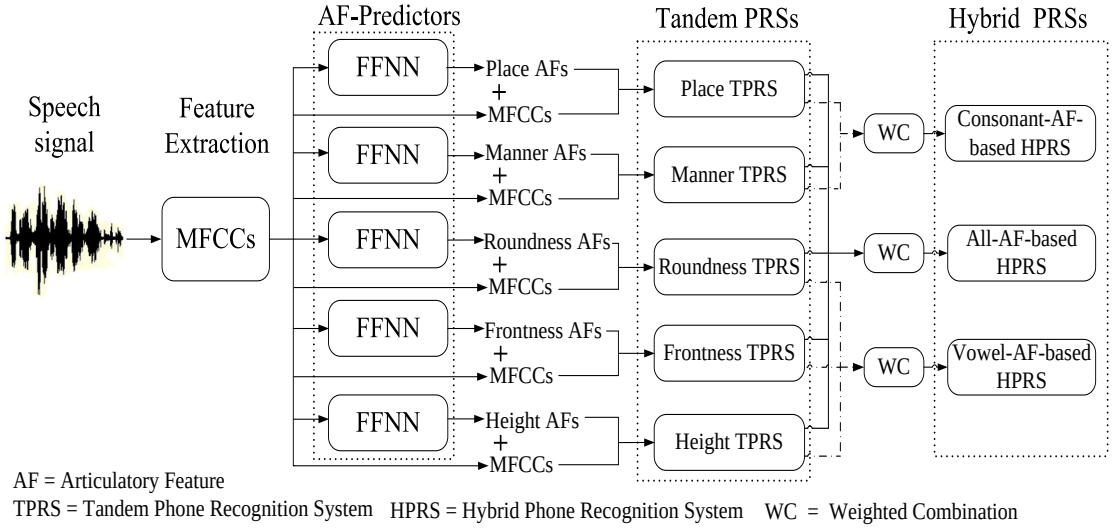


Figure 10.5: Block diagram of Hybrid Phone Recognition Systems.

Table 10.9 shows the optimal weighting factors used for developing hybrid PRSs of read, extempore and conversation modes of speech. The *hyphen* (-) symbol in Table 10.9 indicates that the particular weighting factor is not applicable for the corresponding hybrid PRS. The weighting factors w1, w2, w3, w4 and w5 corresponds to place, manner, roundness, frontness and height AF-based tandem PRSs, respectively.

Table 10.9: Weighting factors used for developing Hybrid PRSs using Weighted Combination Approach.

Hybrid PRS	**Weighting Factors**				
	w1	**w2**	**w3**	**w4**	**w5**
Read					
Consonant-AF-based	0.5	0.5	-	-	-
Vowel-AF-based	-	-	0.3	0.3	0.4
All-AF-based	0.3	0.2	0.2	0.1	0.2
Extempore					
Consonant-AF-based	0.6	0.4	-	-	-
Vowel-AF-based	-	-	0.1	0.4	0.5
All-AF-based	0.3	0.2	0.1	0.1	0.3
Conversation					
Consonant-AF-based	0.5	0.5	-	-	-
Vowel-AF-based	-	-	0.4	0.2	0.4
All-AF-based	0.4	0.1	0.1	0.1	0.3

Table 10.10 shows the phone recognition accuracies of hybrid PRSs. First column lists different types of hybrid PRSs. Second, third and fourth columns show the recognition accuracies of read, extempore and conversation hybrid PRSs, respectively. It is found that the performance of hybrid PRSs

Table 10.10: Phone recognition accuracy (%) of Hybrid PRSs across Read, Extempore and Conversation modes of Speech.

PRSs using different Features	**Recognition Accuracy (%)**		
	Read	**Extempore**	**Conversation**
MFCCs (Baseline)	45.48	39.58	37.20
PP-based Tandem PRS	48.97	40.60	42.14
Consonant-AF-based Hybrid PRS	49.95	43.97	42.05
Vowel-AF-based Hybrid PRS	51.28	44.89	41.52
All-AF-based Hybrid PRS	52.24	45.70	42.97
PP-and-All-AF-based Hybrid PRS	52.61	46.24	44.15

is higher than any of the AF-based tandem PRSs in all the three modes of speech. The improvement in the recognition accuracies of hybrid PRSs is consistent in all three modes of speech. All-AF-based hybrid PRSs have higher recognition accuracy compared to PP-based tandem PRSs. The PP-and-All-

AF-based hybrid PRSs have shown highest recognition accuracy. The highest improvement obtained in the recognition accuracy of read, extempore and conversation modes of speech is 7.13%, 6.66% and 6.95%, respectively. Read speech has higher improvement in recognition accuracy compared to other two modes. The improvement in the performance of conversation speech is nearly same as that of extempore speech.

10.5 Automatic Prosodic Transcription

The automatic prosodic transcription consists of (i) Spotting the syllables in continuous speech, (ii) Transcribing the pitch contour and (iii) Spotting and marking the break indices. Spotting the syllables from continuous speech can be carried out using vowel onset point detection methods. The accuracy of spotting syllables in 3 modes of speech is given in Table 10.11. We have transcribed pitch contour automatically using four labels: very low (VL), low (L), high (H), and very high (VH). We have explored 3 different approaches to transcribe pitch contours: (i) Based on pitch dynamics of each phrase, (ii) Based on pitch dynamics of phrases corresponds to every speaker and (iii) Based on pitch dynamics of all phrases present in speech corpus. Evaluation of proposed automatic pitch transcription has been carried out using 100 sentences of Bengali and Odia in read, extempore and conversation speech modes. Root mean square error between original and automated pitch contours for three modes of speech data are shown in Table 10.12. The break indices are automatically derived using short term energy of the speech signal. Depending on the duration of the detected break, break is transcribed as B1, B2 or B3. It is observed that, accuracy of B2 and B3 detection is better, compared to B1. In this work, B1 refers to intra and inter word breaks, B2 refers to phrase level breaks and B3 refers to sentence level breaks. The details of accuracy of break index transcription is given in Table 10.13.

Table 10.11: Syllable-level segmentation accuracy

	Recognition Accuracy (%)					
	Bengali			Odia		
	Read	Extempore	Conver	Read	Extempore	Conver
Match	82.92	85.31	62.88	72.78	66.03	40.21
Missing	17.07	14.68	37.11	27.2	33.96	59.78
Spurious	10.49	1.07	17.22	12.68	3.63	4.26

Table 10.12: Root mean square error between original and automated pitch contours.

Level		Root mean square error (Hz)					
		Bengali			Odia		
		Read	Extempore	Conver	Read	Extempore	Conver
Phrase		6.01	5.47	17.54	2.91	8.41	8.29
Speaker		6.66	8.60	11.98	4.64	8.53	8.09
Gender	Female	6.22	8.60	15.50	8.79	8.50	8.83
	Male	8.39	5.65	7.66	4.24	3.36	5.48

Table 10.13: Accuracy of Break Indices (B1, B2 and B3)

Break Indices		Recognition Accuracy (%)					
		Bengali			Odia		
		Read	Extempore	Conver	Read	Extempore	Conver
B1	Match	78.12	70.38	58.56	83.76	75.89	62.79
	Missing	21.88	29.62	41.44	16.24	24.11	37.21
	Spurious	61.27	58.43	72.16	52.78	60.13	68.32
B2	Match	85.92	81.31	74.53	88.78	84.03	79.39
	Missing	14.08	18.69	25.47	11.22	15.97	20.61
	Spurious	22.49	26.65	32.58	19.68	25.79	26.57
B3	Match	93.17	88.36	84.67	95.53	90.35	91.48
	Missing	6.83	11.64	15.33	4.47	9.65	8.52
	Spurious	4.28	7.29	8.37	5.78	7.82	6.48

10.6 Search Engine

For automatic retrieval of desired segment of speech, we have explored code book based search engine. Code books of size 32 were derived from vector quantization (VQ). The speech corpus is represented using sequence of code book indices. The sequence of codebook indices are derived from the speech query. The query indices were matched with the sequence of codebook indices of each speech utterance present in the corpus. The matching between the sequences is carried out using VOPs as anchor point. The proposed document search engine performance is analyzed on 455 sentences (1.3 hrs) of Odia read speech corpus. Based on similarity measure, speech utterances are ranked for the given query. The retrieval accuracy is found to be around 30% by considering top 20 documents.

10.7 Publications

Journal

(i) Manjunath K E and K. Sreenivasa Rao, "Source and System Features for Phone Recognition", *International Journal of Speech Technology, (Springer)*, pp. 1–14, 2014.

(ii) Manjunath K E and K. Sreenivasa Rao, "Improvement of Phone Recognition Accuracy using Articulatory Features", *Applied Soft Computing, (Elsevier),* (Under Review).

Conference

(i) S B Sunil Kumar, K. Sreenivasa Rao and Debadatta Pati, "Phonetic and prosodically rich transcribed speech corpus in Indian languages: Bengali and Odia", in *16th IEEE International Oriental COCOSDA (OCOCOSDA-2013)*, (Gurgaon, India), Nov. 2013.

(ii) Manjunath K E, K. Sreenivasa Rao, and Debadatta Pati, "Development of Phonetic Engine for Indian languages : Bengali and Oriya", in *16th IEEE International Oriental COCOSDA (OCOCOSDA-2013)*, (Gurgaon, India), Nov. 2013.

(iii) R Ravi Kiran, Sunil Kumar. S.B, Manjunath K E, Biswajit Satapathy, Apoorv Chaturvedi, Debadatta Pati, and K Sreenivasa Rao, "Automatic Phonetic and Prosodic Transcription for Indian Languages : Bengali and Odia", in *10th International Conference on Natural Language Processing (ICON-2013)*, (New Delhi, India), Dec. 2013.

(iv) Manjunath K E and K. Sreenivasa Rao, "Automatic Phonetic Transcription for Read, Extempore and Conversation Speech for an Indian Language: Bengali", in *20th IEEE National Conference on Communications (NCC-2014)*, (Kanpur, India), Feb. 2014.

(v) Manjunath K E, K. Sreenivasa Rao, and Gurunath Reddy M, "Two-Stage Phone Recognition System using Articulatory and Spectral Features", in *IEEE International Conference on Signal Processing and Communication Engineering Systems (SPACES-2015)*, (Guntur, India), Jan. 2015.

(vi) Manjunath K E, K. Sreenivasa Rao, and Gurunath Reddy M, "Improvement of Phone Recognition Accuracy using Source and System Features", in *IEEE International Conference on Signal*

Processing and Communication Engineering Systems (SPACES-2015), (Guntur, India), Jan. 2015.

(vii) Manjunath K E, Sunil Kumar. S. B, Debadatta Pati, Biswajit Satapathy, and K. Sreenivasa Rao, "Development of consonant-vowel recognition systems for Indian languages : Bengali and Odia", in *10^{th} IEEE India Conference on Emerging Trends and Innovation in Technology (INDICON-2013)*, (Bombay, India), Dec. 2013.