A Neural Network–Based Text–Dependent Speaker Verification System Using Suprasegmental Features

M. Mathew B. Yegnanarayana R. Sundar
Department of Computer Science and Engineering
Indian Institute of Technology, Madras, 600 036, India
E-mail: {mathew@svalpha2, yegna*, sundar}.iitm.ernet.in

Abstract

In this paper, we propose two neural network–based approaches, namely, One Speaker One Network and One Speaker Multiple Networks, for text-dependent speaker verification using suprasegmental features. The suprasegmental features used for this study are pitch accent and durational features. These features are extracted using properties of intonation patterns and duration. We have proposed an approach to combine evidence present at the segmental and suprasegmental levels to improve the performance of the verification system.

1 Introduction

Speaker verification is the process of accepting or rejecting the identity claim of a speaker. The main application of speaker verification is secure access control by voice. The advantage of speaker verification over any other biometric technique, such as finger print identification to recognize a person, is that authentication can be done from a remote place through telephone [1].

When a person speaks, it is difficult to extract speaker-specific characteristics from all phonemes or syllables. The speaker-specific characteristics are present at different levels, namely, segmental level and suprasegmental level. Traditional approaches such as Dynamic Time Warping (DTW) uses segmental feature parameters for speaker verification. The segmental feature parameters are sensitive to transmission channel variations and additive noise. In this paper, we investigate the effectiveness of suprasegmental features for speaker verification. The suprasegmental features are robust against transmission channel variations.

For real applications, it may be useful to combine evidences from different approaches. It is believed that by combining multiple classifiers, the classification performance can be improved [2, 3]. This paper consists of two parts. The first part describes a neural network-based text-dependent speaker verification system using suprasegmental features. In the second part of this paper, we have proposed an approach to combine the evidence present at the segmental and suprasegmental levels within the framework of evidential theory for reliable speaker verification [2, 3].

2 Development of the Speaker Verification System

Speaker verification task consists of three stages: Feature extraction, feature comparison and decision making. It has two operational phases: Training and testing. During training, a speaker-specific model/template is generated from the features extracted from the training data of the speaker. For testing the speaker verification system, the test sample is compared with the references and a decision is taken

*corresponding author

based on the similarity score obtained after comparison.

2.1 Data Collection and Feature Extraction

Speech data from 46 cooperative adult speakers (29 males and 17 females) was collected in text reading style. The speech recording was done in an ordinary office environment. The following sentence in Hindi was selected for developing the system: "mat$\bar{a}$ aur pit$\bar{a}$ k$\bar{a}$ a$\bar{a}$dar karn$\bar{a}$ č$\bar{a}$hiye". Twenty five utterances of the sentence were collected from each speaker. The end-points of the speech signal were determined using the amplitude information of the speech signal.

Suprasegmental features used for this study are pitch accent features and word durations. The F_0 contour is obtained using the properties of the group-delay function [5]. Pitch accent and word duration features are extracted using Word Boundary Hypothesization (WBH) algorithm for continuous speech in Hindi based on F_0 patterns [4]. Fig 1 illustrates the results of feature extraction for an utterance of the test sentence. A 21 dimensional feature vector consisting of twelve pitch frequencies at syllable nuclei and their average pitch, total duration of the utterance and durations of individual words is extracted. The durations are expressed in number of frames.

2.2 Robustness of the Selected Features

The word boundary hypothesization algorithm uses only gross parameters like energy and pitch frequencies in its implementation. Only high SNR portions of the speech signal need to be considered for extracting the pitch accent features. Also the pitch extraction using the group-delay function is robust against noise [5]. Hence the features are robust under noisy input conditions [6]. But still in some cases errors may occur due to errors in finding the peaks of the energy contour, and also due to errors in the pitch contour. These errors may result in wrong placement of the word boundaries. To overcome this problem, a DTW algorithm is used to find the syllable nuclei.

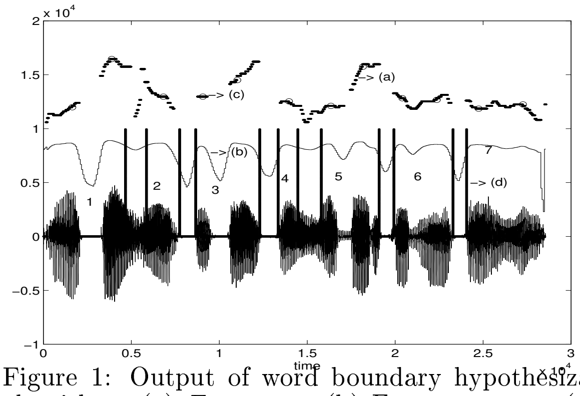

Figure 1: Output of word boundary hypothesization algorithm. (a) F_0 contour (b) Energy contour (c) F_0 at syllable nuclei (d) Word Boundary (1) m$\bar{a}$t$\bar{a}$ (2) aur (3) pit$\bar{a}$ (4) k$\bar{a}$ (5) $\bar{a}$dar (6) karn$\bar{a}$ (7) č$\bar{a}$hiye.

2.3 Neural Network Modeling

Feature vectors of a particular speaker occupies a small region in feature space. For speaker verification, the system has to discriminate between genuine and impostor classes. So there is a need to represent the feature vectors of impostor class. For every speaker, to represent the impostor class, nine background speakers are selected from the corpus such that the speakers average pitch frequency together span the pitch frequency range of 80-400Hz. A two layer feed forward neural network classifier with thirty nonlinear hidden units and two nonlinear output units is used for every speaker [7]. The neural network was trained using ten utterances for each speaker and his background speakers [7]. The speaker verification system was tested with ten utterances of each speaker. The classification decision was based on winner–take–all principle. The results of this One Speaker One Network (OSON) approach are given in Table 1. It was observed that for some speakers the neural network model performed well, whereas for oth-

Table 1: Results of the One Speaker One Network approach

False Acceptance Rate (in %)	False Rejection Rate (in %)	Equal Error Rate (in %)
3.1	10.5	6.8

Table 2: Performance of Individual Networks

Description	False Acceptance Rate (in %)	False Rejection Rate (in %)
21L30N2N	3.1	10.5
21L35N2N	3.5	9.6
21L25N10N	4.7	8.1
21L35N10N	4.6	6.1
21L35N17N10N	5.8	13.0

ers the performance was moderate. The reason for this is attributed to limited training data and variability in the suprasegmental features of these speakers. One approach to solve this problem is to use multiple neural networks to improve the performance. This approach is implemented by training five different neural networks for each speaker. By different neural networks we mean that the architecture and training method is different for different networks, although the same data is used for training. The individual performance of each network in the multiple networks method is given in Table 2.

The output decisions of the networks can be combined using voting theory, i.e, *if three or more neural networks accept the claim, then accept the claim, else reject the claim.* The results of this study are given in Table 3.

Table 3: Results of the One Speaker Multiple Network (OSMN) approach

False Acceptance Rate (in %)	False Rejection Rate (in %)	Equal Error Rate (in %)
2.3	5.7	4.0

3 Combining Evidence using Dempster–Shafer Theory

The multiple classifiers can be combined at measurement level, rank level or abstract level [2]. The voting method suggested in section 2.3 combines the classifiers at abstract level without taking into consideration the confidence with which the input feature vector belongs to the winning class. In this section, we propose an approach to combine the multiple classifiers at measurement level. The units in the output layer of the neural network classifier suggest the support for all the classes. These support values can be combined using Dempster rule for combining evidence [2] to arrive at a decision. In OSMN approach, there are neural network classifiers with two output classes as well as ten output classes. Since the decision in speaker verification is binary, the support information of the classifiers with ten output classes has to be expressed as support for two classes. This can be done in the following way: If the winning unit corresponds to the genuine speaker unit and the support is greater than 0.5 then the support is unaltered, else a support value of 0.6 is assigned to the genuine class. The remaining support out of 1.0 is assigned to the impostor class. The same method is adopted, if an impostor's unit is the winner. For classifiers with two output units, the support for genuine class and imposter class can be estimated by normalizing the output of the corresponding unit by the sum of the output of the two units. These supports are combined using Dempster rule of combining evidence [2]. The results of this study are given in the Table 4.

Table 4: Results of OSMN approach

False Acceptance Rate (in %)	False Rejection Rate (in %)	Equal Error Rate (in %)
1.8	6.0	3.9

3.1 Combining Suprasegmental and Segmental Evidence

In this section, we suggest an approach to combine evidences present at segmental and suprasegmental levels. The segmental level evidence is obtained from a DTW based classifier. The segmental feature parameters are twenty linearly weighted cepstral coefficients and their first five delta coefficients. During the training phase, three reference template for every speaker is generated. Also for every speaker, a genuine DTW distance score distribution and an imposter DTW distance score distribution is generated using the first ten enrollment utterances of genuine and background speakers. During the verification phase, the test utterance is matched with the three reference templates of the claimed speaker. Each DTW match score is converted into a confidence score in favour of genuine class and impostor class by using the genuine class and imposter class distributions, respectively. If the genuine class confidence score for two or all the three DTW matches are more than their respective impostor class confidence score, then the claim is accepted, else the claim is rejected. The first row of Table 5 shows the result of this experiment. The second row of the table shows the result of combining the confidence scores of each match using Dempster rule of combining evidence. The third row shows the result of combining the outputs of the neural networks and the confidence scores of each DTW match using Dempster rule of combining evidences.

Table 5: Results of combining segmental and suprasegmental evidence

False Acceptance Rate (in %)	False Rejection Rate (in %)	Equal Error Rate (in %)
3.6	4.6	4.1
3.6	2.6	3.1
1.0	3.0	2.0

4 Summary and conclusions

The objective of this work was to study the effectiveness of suprasegmental features for speaker verification. The performance of the system based on OSON is affected by variability in the suprasegmental features of speakers, and also due to limited amount of training data. To overcome some of these difficulties, the OSMN approach was proposed where the ability of different neural networks to form different decision surfaces when trained on the same data was exploited. We have shown that by combining the evidence present at the segmental and suprasegmental levels the performance of the system can be improved significantly.

References

[1] D. O'Shaughnessy, " Speaker Recognition," *IEEE ASSP Magazine*, pp. 4–17, Oct. 1986.

[2] L. Xu and A. Kryzak and C. Y. Suen, "Methods of combining multiple classifiers and their application to handwriting recognition," *IEEE Trans. Systems, Man and Cybernetics*, vol. SMC–22, pp. 418–435, May. 1992.

[3] K. Chen and L. Wang and H. Chi, "Methods of combining multiple classifiers with different features and their application to text-independent speaker identification," *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 11, no. 3, pp. 417–445, 1997.

[4] S. Rajendran and B. Yegnanarayana, "Word boundary hypothesization based on F_0 pattern," *Speech Communication*, vol. 18, pp. 21–46, 1996.

[5] B. Yegnanarayana and V. R. Ramchandran, "Group delay processing of speech signals," *Proc. ESCA Workshop on a Comparing Speech Signal Representation*, pp. 411–418, Sheffild, England, 1992.

[6] B. Yegnanarayana and S. P. Wagh and S. Rajendran, " A speaker verification system using prosodic features," *Proc. ICSLP*, pp. 1867–1870, Yokohama, Japan, 1994.

[7] B. Yegnanarayana, *Artificial neural Networks*, Prentice Hall of India Pvt. Limited, New Delhi, 1998