PROCESSING THROAT MICROPHONE SPEECH

A THESIS

submitted by

A. SHAHINA

for the award of the degree

of

DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY MADRAS

CHENNAI - 600036

MAY 2007

To

my parents, Abdul Shukur and Shahanaz

and

Nayeem, Safiyyah and Aqeel

THESIS CERTIFICATE

This is to certify that the thesis entitled **Processing Throat Microphone Speech** submitted by **A. Shahina** to the Indian Institute of Technology, Madras for the award of the degree of Doctor of Philosophy is a bonafide record of research work carried out by her under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Madras - 600 036 (Prof. B. Yegnanarayana) (Dr. C. Chandra Sekhar)

Date:

ACKNOWLEDGEMENTS

Praise be to the Lord of the worlds. He has blessed me with the wonderful oppurtunity of pursuing my Ph.D degree at Indian Institute of Technology-Madras under the guidance of the illustrous Professor B. Yegnanarayana. He stands tall - a no-nonsense attitude, excellent teaching, an incredible passion for research, ability to multitask N research topics of N students, modesty, benevolence, constantly encouraging and driving his students towards their research goals - he is akin to the master craftsman who can produce the finest of masterpieces with his ingenuity. His attitude and approach have moulded me into a better version of myself. I shall remain ever grateful to him for having taken me under his tutelage.

I am indebted to Dr. C. Chandra Sekhar for having taken me under his guidance in the final stages of my course. I have benefited from his numerous suggestions and ideas during the various technical discussions. He has been very kind in ensuring that I pursue my research and enjoy all the facilities in the lab, unhindered. I also thank him for meticulously correcting my thesis draft, and for the enormous patience he has shown towards me.

I profoundly thank Dr. S. Rajendran for evincing keen interest in some of my research problems, and for his prompt help on several occasions. I am thankful to Dr. Hema A. Murthy for her concern, encouragement and thought-provoking questions during various technical presentations.

I thank Prof. S. Raman and Prof. T.A. Gonzalvez, Chairpersons of the department of Computer Science and Engineering, for providing me constant support and excellent facilities to carry out my research. I am also thankful to my doctoral committee members, Prof. Megha Singh, Prof. K. M. M. Prabhu, Prof. C. Siva Ram Murthy, Prof. S. V. Raghavan, and Prof. S. Mohan for their useful advice, and for sparing their valuable time to evaluate the progress of my research work. I am grateful to all the faculty members of the department for imparting knowledge in various subjects. I thank the staff of the office of computer science and engineering department, Mr. Natarajan, Ms. Saradha and Mr. Imakaran, and the staff of Industrial Consultancy and Sponsored Research (ICSR) for their help in administrative matters.

I am grateful to the Center for Artificial Intelligence and Robotics (CAIR), Bangalore for having sponsored my research. The formal and informal interactions with the scientists at CAIR, especially the workshops, were very inspiring. I want to specially thank Mr. M. R. Kesheorey, senior scientist at CAIR, for encouraging me and being supportive throughout my research work.

My research journey was made easier on many occasions, thanks to the wonderful team of research scholars, both past and present, in the speech and vision lab. I thank every one of them - Mahadeva Prasanna, Sreenivas Rao, Palanivel, Suryakanth Gangashetty, Anil Kumar Sao, Sri Ram Murthy, Guruprasad, Dhananjaya, Anand Joseph, Leena Mary, Kumaraswamy, Krishnamohan, Satish, Chaitanya, Suresh, Swapna, Chandrakala, Sheetal, Panuku, Ved Prakash, Anita, Harnadh, Vijay Ram Raju and Harendra. I am especially indebted to Guruprasad, Sri Ram Murthy and Dhananjay, who have helped me a lot with their suggestions and critical feedback.

I am ever indebted to all my family members. My ever-loving parents, Mr. A. Abdul Shukur and Mrs. A. Shahanaz, with their consistent prayers, sacrifice, patience, unwavering support and encouragement, helped me sail through. My husband, A. Nayeemulla Khan has been my moral strength and support throughout, ever-willing to ease my burdens. Our blossoming flowers, N. Safiyyah and N. Muhammad Aqeel, spread the fragrance of joy (which, God willing, shall not wither) which overwhelms all the stresses and strains. Glory be God. Everything is as He wills.

A. Shahina

ABSTRACT

The speech from a skin vibration transducer, called the throat microphone, is intelligible, but sounds unnatural. The Throat Microphone (TM) picks up speech that is transmitted from the pharynx region, and the 'buzz tone' of the larynx. The main objective of this thesis is to improve the naturalness of the TM speech.

A study of the acoustic characteristics of various sound units in the TM and Normal Microphone (NM) speech shows that the TM and NM signals differ in the vocal tract characteristics as well as in the characteristics of the excitation source for different sound units. However, though there are acoustical differences, there exist some common features (for example, pitch and location of formants) in the simultaneously recorded TM and NM speech of a speaker. This speaker-specific correlation can be exploited to learn a mapping between the features of the TM and NM speech, so as to improve the perceptual quality of the TM speech.

MultiLayered FeedForward Neural Network (MLFFNN) is used to obtain a smooth mapping of the TM spectrum onto the NM spectrum for each frame, without 'spectral jumps' between adjacent frames. The all-pole filter derived from the spectral mapping is guaranteed to be stable, because the LP coefficients are obtained using the autocorrelation coefficients derived from the mapped cepstral coefficients. The excitation source information in the TM and NM speech differ in the relative strengths of the instants in the voiced sound units. While in the NM residual the strength of the in-

stants is comparatively high for vowels and low for voiced stops, in the TM residual the strengths are comparable for all the voiced sounds. Hence modification of the TM residual involves mapping features, that discriminate at least among the broad voiced sound categories, from the TM speech onto the NM speech. The parameters derived from the mapping are used to emphasize the strength of the instants in the vowel regions and deemphasize the instants in the voiced consonant regions. The modified TM residual is used to excite the all-pole synthesis filter derived from the spectral mapping to obtain the enhanced speech.

One of the advantages of using a throat microphone is that it provides a high Signal-to-Noise Ratio (SNR) over a large part (from 0 to 3500 Hz) of the audio frequency range. This thesis explores the presence of speech, speaker and language characteristics in the TM speech for developing speech systems. An HMM based syllable recognizer is developed using the TM speech to recognize the 145 consonant-vowel units of Hindi. There exist many situations which require robust speech systems whose performance does not degrade due to ambient noise. For example, restricted entry into high security enclosures, and handling calls from non-native speakers in noisy environments. As the TM speech is relatively immune to noise, the TM speech is used to build robust speaker recognition and language identification systems.

In recent years, there has been considerable interest in improving the perceptual quality of the narrowband (300-3400 Hz) telephone speech. The approach used to improve the naturalness of the TM speech is extended for bandwidth extension of the narrowband telephone speech and loudness enhancement of soft voices.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AANN	- AutoAssociative Neural Network
ARMA	- Auto Regression and Moving Average
CV	- Consonant-Vowel
CVC	- Consonant-Vowel-Consonant
GMM	- Gaussian-Mixture Model
HE	- Hilbert Envelope
HMM	- Hidden Markov Model
LID	- Language IDentification
LSF	- Line Spectral Frequency
LP	- Linear Prediction
LPC	- Linear Prediction Coefficients
MFCC	- Mel Frequency Cepstral Coefficients
MLFFNN	- MultiLayered FeedForward Neural Network
MOA	- Manner Of Articulation
MMSE	- Minimum Mean Square Error
NM	- Normal Microphone
NN	- Neural Network
POA	- Place Of Articulation
SPLICE	- Stereo-based Piecewise Linear Compensation for Environments
SVM	- Support Vector Machine
TM	- Throat Microphone

UVA	- Unvoiced Aspirated
UVUA	- Unvoiced Unaspirated
VA	- Voiced Aspirated
VC	- Vowel-Consonant
VCV	- Vowel-Consonant-Vowel
VUA	- Voiced Unaspirated
wLPCCs	- weighted Linear Prediction Cepstral Coefficients

NOTATION

Lower case boldface letters are used to denote vectors and upper case boldface letters to denote matrices. In addition, the following convention is used throughout the thesis:

n	– discrete time index
p	– linear prediction order
a_n	– linear prediction coefficient
q	– dimension of weighted linear prediction cepstral coefficients
c_n	– linear prediction cepstral coefficient
$H(k)$	– linear prediction spectrum
$S(k)$	– linear prediction log spectrum
$P(k)$	– estimate of linear prediction spectrum obtained from spectral mapping
$R(n)$	– autocorrelation function
E	– mean square error
η	– normalized prediction error
R_η	– ratio of normalized prediction errors of TM speech and NM speech
G	– log frame energy
$s(n)$	– speech signal
$r(n)$	– linear prediction residual signal
$h(n)$	– Hilbert envelope of residual signal
$r_h(n)$	– Hilbert transform

$r_a(n)$	– analytic signal corresponding to residual signal
$cos(\theta(n))$	– cosine of the phase of the analytic signal
$A(z)$	– inverse filter

CHAPTER 1

PROCESSING THROAT MICROPHONE SPEECH: AN INTRODUCTION

1.1 OBJECTIVES OF THE THESIS

Speech signal recorded from a skin vibration transducer placed near the larynx, called the throat microphone, is intelligible, but sounds unnatural. As humans are accustomed to speech that radiates from the mouth (as recorded by a normal microphone), they are not comfortable when subjected to prolonged hearing of speech from a throat microphone. The main objective of this thesis is to improve the perceptual quality of the throat microphone speech, so as to make it sound natural.

One of the advantages of a throat microphone is that it provides a good Signal-to-Noise Ratio (SNR) over a large part (from 0 to 3500 Hz) of the audio frequency range, which is required for good intelligibility. This thesis explores the presence of speech, speaker and language characteristics in the throat microphone speech for developing speech systems which may be useful in situations where normal microphones cannot be used.

There has been considerable interest in providing high quality speech to subscribers of the telephone channel. The telephone speech is band-limited (300-3400 Hz), and hence its perceptual quality is reduced compared to the speech from a normal mi-

crophone. The technique used for improving the perceptual quality of the throat microphone speech is extended to improve the bandwidth of the telephone speech, and enhance the loudness of soft voices.

1.2 SPEECH PRODUCTION

Speech is produced as a result of time-varying excitation of a time-varying vocal tract system. The air expelled from the lungs causes the vocal cords to vibrate. This vibration at the glottis causes quasi-periodic pulses of airflow into the region above the glottis (supraglottal region). The supraglottal region consists of three main areas: the pharynx, the oral cavity and the nasal cavity. The pharynx consists of the area above the larynx and below the uvula (the flesh blob that hangs down in the back of the throat), as shown in Fig. 1.1. The quasi-periodic pulses of airflow are modified by the articulators (like tongue, teeth, lips etc.) as they pass through the pharynx and the oral cavity. The airflow is also modified through the path of the naso-pharynx, when the velum is lowered, and through the nostrils. The sound waves radiated at the lips and the nostrils are the result of the modifications imposed on the pulses of air flow by this entire resonating system.

Voiced sounds are produced when the source of excitation is the vibrating vocal folds. Unvoiced sounds are produced when air is forced to flow through a narrow constriction in the vocal tract, giving rise to a turbulence. The turbulence results in random movement of air particles in contrast to the quasi-periodic flow of air in the voiced sounds. The noise source for the unvoiced sounds may be located at different places in the vocal tract corresponding to the place of articulation of the sound. The

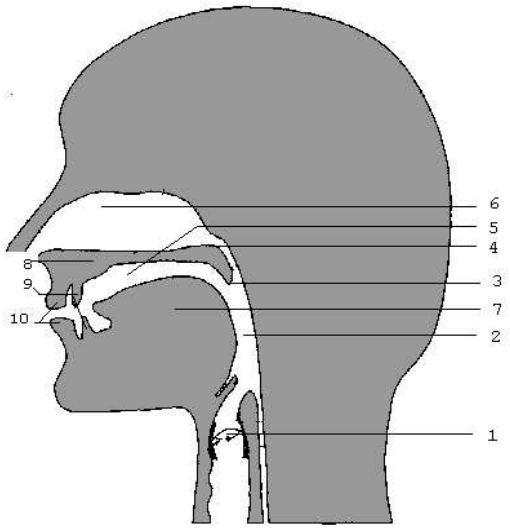

Figure 1.1: The supraglottal region of the speech production mechanism: 1-glottis, 2-pharynx, 3-uvula, 4-velum, 5-oral cavity, 6-nasal cavity, 7-tongue, 8-alveolar ridge, 9-teeth, 10-lips.

location of the noise generator affects the dimensions of the section of the vocal tract. The sections of the vocal tract, both behind and in front of the constriction, play their part in determining how the noise will be modified by the system. The noticeable effect of the vocal tract is the drastic reduction in the amplitude of noise energy over certain frequency bands (filtering effect of the vocal tract system). This can be observed in the case of fricative sounds. In the case of stop consonants, the sudden release after the closure results in a very short burst of noise. This is again modified by the vocal tract filter depending on the place of articulation of the sounds. The speech production mechanism is capable of infinite modifications owing to the flexibility of the articulators.

The speech radiated through the lips and nostrils is picked up by the conventional microphone (referred to as the normal microphone in this thesis) placed in front of the

mouth. The significance of the speech recorded using a normal microphone is that it contains the entire frequency spectrum of the various speech components. Due to high intelligibility and perceptual quality of the signal, the normal microphone is widely used to record speech signals.

Speech signals are not only radiated through the lips and nostrils, but are also propagated through the vibrations of the body tissue. These vibrations are picked up by alternate speech sensors. The pick-up of intelligible speech sounds from anatomical vibrations is not limited in location. Some of the locations are the throat skin, the facial skin (cheek), the underside of the chin and the skull bone behind the ear. The sensors are placed in contact with the skin or bone to pick-up the vibrations. The throat microphone is one such skin vibration transducer.

1.3 THROAT MICROPHONE

The throat microphone device used in this study comprises of a pair of moulded housings mounted on a neckband. One housing is fitted with a microphone transducer, while the other is a dummy unit (refer Fig. 1.2). The twin housings are located on each side of the throat. Soft leather pads are provided to give low contact pressure. As a result, wearer comfort is maintained during long periods of use. A moving member or 'button' transmits the throat vibration mechanically to an outer diaphragm. The vibrations of the outer diaphragm are transmitted acoustically to an inner diaphragm coupled to the transducer. The transducer is a high quality moving iron magnetic type [1].

The throat microphone is placed in contact with the skin slightly above the larynx

Figure 1.2: (left) A throat microphone, and (right) a person wearing the throat microphone.

such that it does not restrict the head motion of the person wearing it (refer Fig. 1.2). It picks up intelligible speech sounds transmitted from the pharynx through the throat tissue. Due to proximity to the larynx, the buzz tone of the larynx is also picked up by the throat microphone, but is overbalanced by the speech sounds. Most of the voiced sounds are easily understood, as the source of excitation of these sounds is the larynx. However, unlike the Normal Microphone (NM) speech which results from the modification in the oral and nasal cavities, the Throat Microphone (TM) speech may not pick up the finer articulatory modifications in the oral cavity. Additionally, the damping effect of the nasal cavity may not be well represented in the TM speech. In the case of unvoiced sounds like fricatives, the throat microphone picks up the noise-like signal from the cavity behind the constriction, unlike the normal microphone which picks up the signal after it passes through the oral cavity in front of the constriction. The effect of the back cavity on the TM speech is to filter out most of the higher

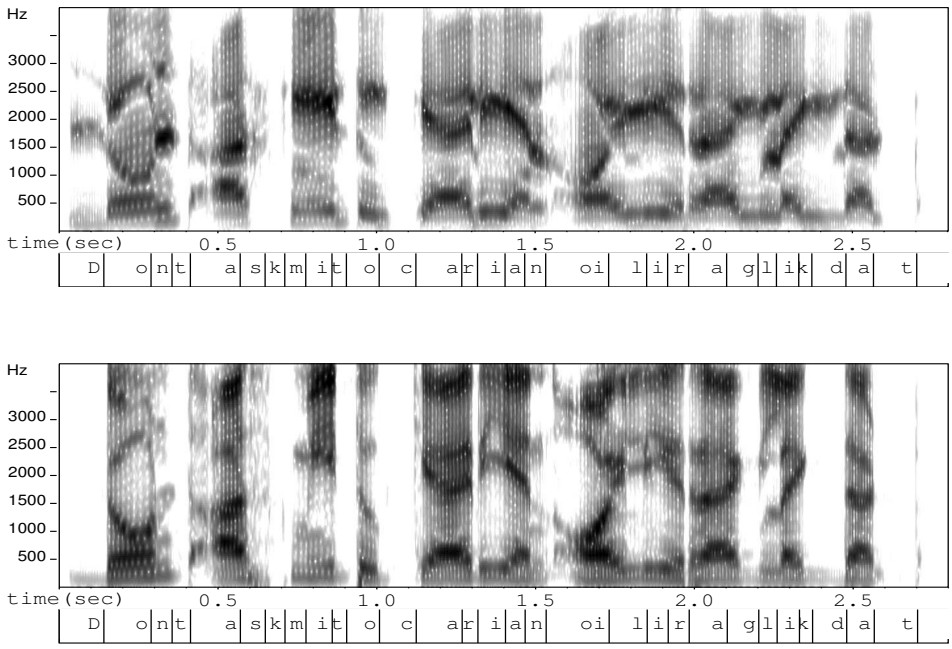

Figure 1.3: The wideband spectrograms of a speech signal from a male speaker recorded simultaneously from (a) a throat microphone, and (b) a normal microphone, for the sentence *don't ask me to carry an oily rag like that.*

frequency bands in the fricatives. In the case of voiced stops, the TM captures the voicing activity during the closure phase as well as the modifications of the air flow in the oral segment behind the closure. In contrast, the NM picks up only the low frequency vibrations of the throat since the NM is located in front of the closure.

Fig. 1.3 shows the spectrograms of a segment of speech simultaneously recorded using a throat microphone and a normal microphone. It is seen from the figure that some of the higher frequencies are missing (e.g., fricative */s/*), or are of low intensity (e.g., formants around 3500 Hz) in the TM speech. Also, additional spectral information is seen in some sounds like */D/* in the TM speech. The TM and NM are sensitive

to different aspects of the speech signal, and their spectra differ as a function of the speaker, the placement of the sensors and as a function of the articulation of speech itself. However, most of the information in the speech components (up to 3000-3500Hz) in the audio frequency band is well preserved in the TM speech, which is required for intelligibility of the speech. Subjective tests on the intelligibility of the TM speech is discussed in the following subsection.

1.3.1 Intelligibility of TM speech

Intelligibility refers to the percentage of the message in the speech that is successfully transmitted from the speaker to the listener. Intelligibility is one important aspect of the overall speech quality [2]. The intelligibility of the TM speech is tested to measure

- the intelligibility of consonant phonemes
- the intelligibility of vowels
- contextual intelligibility

The test utterances of five speakers are evaluated by 10 listeners (different from the speakers).

The intelligibility of consonant phonemes is measured using the Diagnostic Rhyme Test (DRT). In the DRT, a closed set response test, each test is restricted to a pairwise comparison [2, 3]. Each pair of rhyming words differ only in the leading consonant phoneme. The listeners have to indicate which of the two words they heard. Six intelligibility attributes are utilized by the DRT, namely, voicing, nasality, sustention, sibilance, graveness and compactness. A total of 192 words (16 word pairs in each of

the 6 categories) are recorded using the throat microphone from five speakers. The speech data for each category comprises of words spoken by all the speakers.

In the voicing category (phonemes produced by the vibration of the vocal folds, e.g., *Dint* vs. *Tint*), all the phonemes are correctly identified by the listeners. This indicates that all the voiced oral sounds are picked up well by the throat microphone. In the nasality category (phonemes produced through the nasal radiation, e.g., *News* vs. *Dues*) too, the words are generally correctly identified, though misidentification occurs when there is a confusion between */m/* and */b/*, both having the same place of articulation (bilabial sounds). This indicates that though the damping in the nostrils is not well picked up by the TM, generally the nasal sounds are identified. This could be because the TM captures the modifications in the naso-pharyngeal tube. Misidentification occurs in the sustention category (phonemes produced by a partial closure of the vocal tract such as */sh/* or */v/* as against the phonemes produced by complete closure of the vocal tract, e.g., *Sheet* vs. *Cheat*). This could be because the throat microphone is not able to capture the difference in the turbulence generated in the oral cavity during the articulation of a fricative or the release of an unvoiced stop. In the sibilance category (affricated phonemes e.g., *Jaws* vs. *Gauze*), misidentification occurs due to confusion between */c/* and */k/*, and */j/* and */g/*. However, the words are generally correctly identified. In the graveness category (bilabial sounds as against alveolar or dental sounds e.g., *Pool* vs. *Tool*) and compactness category (velar and palatal sounds as against other sounds e.g., *Key* vs. *Tea*), the words are generally correctly identified, though misidentification occurs due to confusion between */s/* and */sh/*.

Though the consonants are generally identified correctly, misidentification occurs as the turbulence due to the articulation of fricatives and release of stop sounds is not well differentiated in the TM speech. The overall DRT score (refer Table 1.1) obtained by the correct percentage of the response, is given by

$$DRT\% = \frac{N_c - Ni}{N_t} \times 100, \tag{1.1}$$

where N_c is the number of correct responses, N_i is the number of incorrect responses, and N_t is the number of tests conducted.

The intelligibility of vowels in the TM speech is measured using the "sustained vowel list" [3]. Each of the three lists in the test consists of 14 Consonant-Vowel-Consonant (CVC) words, such that each of the fourteen vowel phonemes of general American English is represented in the medial vowel of each of the words in each list. This test is useful for isolating the vowel sounds as opposed to the consonant sounds as a measure of intelligibility. The words are correctly identified by most of the speakers (refer Table 1.1). This shows that the vowels are intelligible in the TM speech.

The contextual intelligibility of the TM speech is measured using the Sentence List Tests. The Harvard Psychoacoustic Sentences, which are phonetically balanced, are used for this test. There are 72 sets of 10 meaningful sentences each. Two sets have been used in this test. The result of the intelligibility tests (shown in Table 1.1) shows that the contextual intelligibility is high in the TM speech.

Humans are trained to listen to speech radiated from the lips and the nostrils. Hence, though the TM speech is intelligible, it is perceived as unnatural. Humans are uncomfortable when subjected to prolonged hearing of speech from a throat mi-

Table 1.1: The scores (in %) of the intelligibility tests performed on the TM speech recorded in a clean environment

Speaker	Word list DRT test	Sustained vowel list	Sentence list Har. Psyc. test
1	93	98	96
2	83	91	90
3	89	93	88
4	80	87	86
5	83	90	87

crophone. It becomes necessary to improve the perceptual quality of the TM speech to alleviate the discomfort of the listeners. From the discussions in this section, it is seen that though there are acoustic differences between the TM and NM speech for various sound units, there exists correlation in the simultaneously recorded TM and NM speech signals of a speaker. This is because the information about the vocal cord activity (e.g., pitch) and the dimensions of the vocal tract (e.g., location of formants) of a speaker remain similar in both the TM and NM speech. This speaker-specific correlation can be exploited to learn a mapping between the features of the TM and NM speech. This could improve the perceptual quality of the TM speech.

1.3.2 Robustness of TM speech

The throat microphone is a preferred choice for use in situations that demand an intelligible speech with a high SNR in noisy ambience, and where physical comfort

and convenience of the user is essential to ensure that his/her freedom of action is not limited. They are also used when operational constraints prevent the use of a normal microphone.

Typical applications, where high noise causes communication difficulties, include helicopters or fixed wing aircraft, armoured vehicles, naval vessels and severe industrial conditions. Throat microphones are also useful where mechanical compatibility with respirator masks or military helmets prevents the use of normal microphones. Hence, it is worn by safety personnel in extreme conditions like raging forest fires or driving in hurricanes and by security personnel in the midst of gunfire. It can also be used in chaotic places like railway stations, and by cellphone users. Fig. 1.4 shows the spectrograms of speech simultaneously recorded using a TM and an NM in (simulated) noisy environment. It is seen that the TM speech has a high SNR, while the NM speech has a low SNR.

There exist many situations which require robust speech systems whose performance does not degrade due to the ambient noise. Restricted entry into high security enclosures, access control, and command and control are some of the typical applications. The presence of speech, speaker and language characteristics in the robust TM speech can be explored for developing reliable speech systems, especially in adverse conditions.

1.4 BANDWIDTH EXTENSION OF TELEPHONE SPEECH

In recent years, significant efforts have been made to improve the quality of telephone speech signals. The speech signals from the analog telephone channels are band-limited

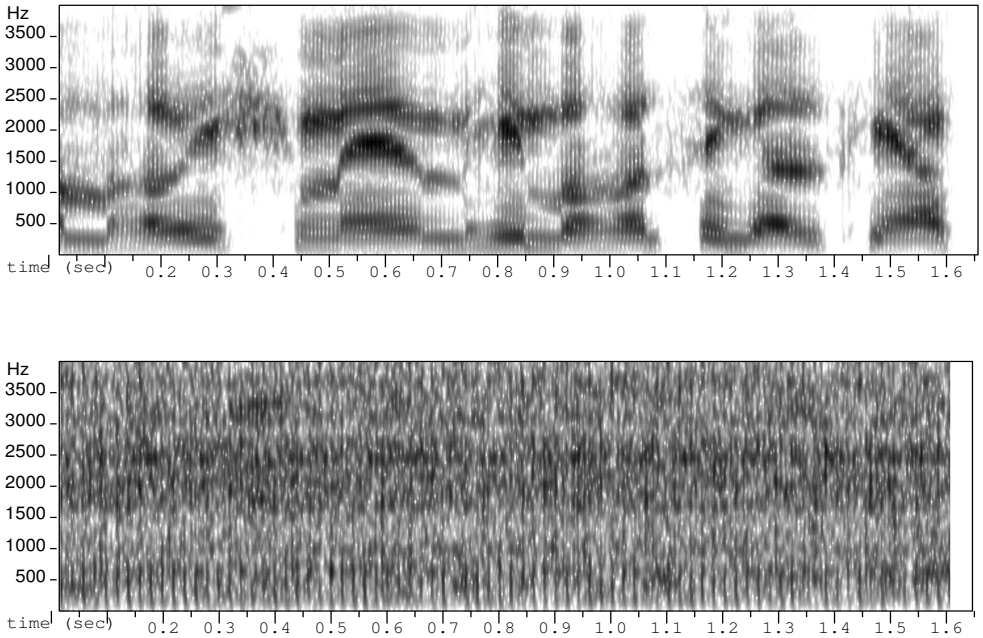

Figure 1.4: The wideband spectrograms of a speech signal from a male speaker recorded under simulated noisy condition, simultaneously from (a) a throat microphone, and (b) a normal microphone.

between 300 Hz and 3400 Hz. This limitation in the audio frequency range is caused by the introduction of band-limiting filters within amplifiers used to keep a certain signal level in long local loops [4]. These filters have a passband from approximately 300 Hz up to 3400 Hz, and are applied to reduce crosstalk between different channels. As the higher frequencies are missing in the narrowband telephone speech, sounds that have predominant energy in the higher frequencies (as in fricatives) are weak. Also, some of the stop consonants are not easily distinguishable. These are some of the reasons that reduce the perceptual quality of the telephone speech, though it is intelligible. The efforts to improve the perceptual quality of the telephone speech involve estimating

the speech signal components above 3400 Hz, and complementing the signal in the idle frequency bands with this estimate. The estimation of the higher frequency bands exploits the redundant information in the lower frequency bands of the narrowband telephone speech and the wideband normal speech (0 to 8000 Hz). The approach to enhance the TM speech, which exploits the redundant information in the TM and NM speech (refer Section 1.3.1), can be used to increase the bandwidth of the telephone speech.

Techniques used in this work for enhancing the TM speech are extended for two applications, (1) bandwidth extension of telephone speech, and (2) loudness enhancement of soft voices.

1.5 SCOPE OF THE PRESENT WORK

The primary problem addressed in the thesis is to improve the naturalness of the TM speech, close to that of the speech recorded from a normal microphone. There exists correlation in the TM and NM speech of a speaker. The activity of the vocal folds and the dimensions of the vocal tract are specific to a speaker. So, in general, the locations of instants of significant excitation and the lower formant locations would be similar in the simultaneously recorded TM and NM speech from a speaker. This thesis exploits the correlation to learn a speaker-specific mapping from the features of the TM speech to the features of the NM speech. The estimated features of the NM speech are used to reconstruct the enhanced speech. The recordings for learning the mapping and testing are done in laboratory conditions.

In building speech systems using TM speech in noisy environments, the perfor-

mance of the speech systems depends critically on the effect of the environmental conditions on the parameters and features extracted from the speech signal [5]. The quality of the TM speech is not affected in the presence of noise such as running of engines, reverberation and competing speaker. However, in the presence of severe vibratory noise, the SNR of the TM speech reduces. This is because the TM picks up the vibrations in the throat skin due to the vibratory noise. In this thesis, the noisy environment is simulated in the laboratory using the radio statics (refer to Fig. 1.4). This noise does not affect the quality of the TM speech.

1.6 ORGANISATION OF THE THESIS

The focus of the work presented in this thesis is on processing the TM speech for (a) improving the naturalness and (b) building robust speech systems in noisy conditions. Techniques for improving the naturalness of the TM speech have been extended to improve the perceptual quality of the narrowband telephone speech. The evolution of ideas presented in this thesis is given briefly in Table 1.2.

The contents of the thesis are organised as follows:

Chapter 2 reviews some of the alternate speech sensors and the methods used to process the signals recorded from them. Methods to improve the quality of telephone speech are also discussed.

In Chapter 3, a study of the acoustic characteristics of various sound units in the TM speech in comparison with the characteristics of the sound units in the NM speech is presented. The sound units studied are vowels, nasal consonants, stop consonants, fricatives and semivowels.

The methods to enhance the naturalness of the TM speech are presented in Chapters 4 and 5. The spectral based enhancement is presented in Chapter 4, while the excitation based enhancement is presented in Chapter 5.

In Chapter 6, speech recognition studies using the TM speech are presented. A syllable recognizer is developed for isolated utterances in an Indian language, Hindi, using the TM speech in clean conditions.

In Chapter 7, two robust speech systems using TM speech recorded in (simulated) noisy conditions are presented. The speaker recognition and language identification systems perform reliably in clean as well as noisy conditions.

In Chapter 8, the techniques to improve the naturalness of the TM speech have been used to improve the quality of the narrowband telephone speech. A similar technique is shown to be useful in enhancing the perceptual loudness of any soft voices that are recorded using a normal microphone.

Chapter 9 summarises the work presented in this thesis. The contributions of this research are highlighted and some directions for future work are given.

Table 1.2: Evolution of ideas presented in the thesis.

- The TM speech is perceptually unnatural compared to the NM speech
  - Acoustic analysis of various sound units shows TM and NM speech differ in spectral content and excitation source characteristics of voiced sounds.
- Approaches to improve perceptual quality of TM speech
  - Mapping TM spectra onto NM spectra using neural networks
    - Continuous mapping to ensure synthesized speech does not suffer from spectral discontinuities
    - Method to guarantee stability of the synthesis filter
  - Modifying TM residual
    - Mapping features that distinguish between voiced segments in TM speech and NM speech
    - Emphasizing strength of instants in vowels, and deemphasizing instants in voiced consonants using mapped features
- Throat microphone is preferred in adverse situations since the speech is intelligible
  - HMM based syllable recognizer developed using TM speech recorded in clean conditions
  - Robust speaker recognition and language identification systems developed using TM speech recorded in (simulated) noisy conditions
- Significant efforts made in recent years to improve perceptual quality of narrowband telephone speech
  - Spectral mapping technique and residual modification technique used for bandwidth extension of narrowband telephone speech and loudness enhancement of soft voices

CHAPTER 2

REVIEW OF APPROACHES TO PROCESS SIGNALS FROM ALTERNATE SPEECH SENSORS

In this chapter, a review of the methods for processing speech signals from various sensors is presented. The chapter is organized as follows. Section 2.1 gives an overview of processing speech signals from a standard microphone. Section 2.2 reviews the methods to process signals from alternate speech sensors, especially for improving the performance of different speech systems. Section 2.3 gives an overview of the methods to improve the perceptual quality of the narrowband telephone speech. Section 2.4 presents an overview of the thesis. Section 2.5 summarizes the reviewed work.

2.1 PROCESSING SPEECH SIGNALS

Speech signals recorded from a normal microphone (placed in front of the speaker's mouth) contain a wide range of audio frequencies. Hence, this speech sounds perceptually natural. This Normal Microphone (NM) speech signal carries information about the speech, speaker and the language spoken, among others. Features containing this information are extracted from the NM speech for several applications such as automatic speech recognition, speaker recognition/verification and language identification. The features are obtained by processing the NM speech signal to estimate the time-varying vocal tract system characteristics and the excitation source characteristics of the speech production mechanism [5–11].

However, in adverse situations like high ambient noise, the speech degrades and hence the performance of the speech systems. These NM speech based systems suffer from limitations such as requiring estimates of the speech spectrum and speech activity detection from a noisy acoustic waveform. Approaches to improve the performance of these speech systems by enhancing the degraded speech as a preprocessing step have been attempted. Alternate approaches to improve the performance of the speech systems have emerged that capitalize on recent developments in alternate speech sensors. These sensors are relatively immune to the acoustic background noise, and thus provide the potential for robust measurement of speech characteristics. Some of the alternate speech sensors and the approaches that use these sensors for improving the performance of speech systems are discussed in the following section.

2.2 PROCESSING SPEECH SIGNALS FROM ALTERNATE SPEECH SENSORS

Alternate speech sensors including skin vibration, bone conduction, and microwave radar sensors have recently been applied to the problem of measuring speech signals in the presence of strong background noise. Some sensors record speech from anatomical vibrations. Some of the sensors provide measurements of functions of the glottal excitation, and of the vocal tract articulator movements that are relatively immune to the acoustic disturbances, and can supplement the acoustic speech waveform. These sensors have been historically used almost exclusively in clinical environments for applications such as pitch determination [12]. However, recent research has begun to consider this potential for improving speech quality in highly noisy environments. Sensors that typically measure some information about the speech generation process are

commonly used in conjunction with a normal microphone and addition signal processing, to augment the acoustic speech signal and to improve the resulting speech quality. Some of the alternate speech sensors that have been used for speech applications are discussed in the following section.

2.2.1 Alternate speech sensors and measurements

The ElectroGlottoGram (EGG) is a device designed to measure the glottal activity (contact between the vocal folds). The EGG nominally measures the vocal fold contact area (VFCA). The electrodes are placed on the subject's neck at the level of the thyroid cartilage. The VFCA is measured by observing the variation in electrical impedance over time. However, EGG does not sense the events during the open phase of the glottis [13].

The Glottal Electromagnetic Micropower Sensor (GEMS) is a device based upon transmitting ElectroMagnetic (EM) waves into the glottal region [14]. A small antenna is placed on or near the throat at the level of the glottis. This antenna transmits a 2.3 to 2.4 GHz low power EM wave. The reflected signal depends on the tissue movement in the speech production anatomy, such as the tracheal wall, the vocal folds, or the vocal tract wall. The GEMS sensor is able to detect the transition boundaries between voiced and unvoiced or no speech. It measures quasi-periodic signals during the production of vowels, nasals and voiced stops (during closure phase).

The Tuned Electromagnetic Resonator Collar (TREC) sensor is designed to measure glottal activity like the EGG and the GEMS sensors. In measures changes to the relative permittivity of the larynx as a proxy for measuring movement of the glottis (as

well as movement of tissue in the subglottal and supraglottal systems). The relative permittivity is measured by placing capacitors on a collar worn around the speaker's neck, without making contact to the skin [3, 12].

The Physiological microphone (P-mic) is typically placed in the throat area below the glottis. It is composed of a gel-filled chamber and piezo-electric sensor behind the chamber. The sensor converts the vibrations permeating the chamber into electrical signals. When placed below the larynx, it measures the vocal fold vibrations. When strapped to a facial location (forehead) it gives information about the vocal tract such as the noise component of unvoiced speech, but the SNR of the signal reduces [13]. In contrast, the throat microphone gives significant vocal tract information when placed at the level of the throat, as discussed in Chapter 1.

The bone-conduction microphone is a sensor that touches the side of a person's face directly in front of the ear. It receives vibrations that are conducted from the rear of the vocal tract through the flesh and bone of the face when a person speaks. Like the P-mic, the bone-conduction microphone comprises a piezo-electric material. When placed at the skull location, the bone-conduction microphone is found to provide strong voicing, as well as significant vocal tract information in the low frequency range (less than 3 kHz) [15–17].

The silicon Non-Audible Murmur (NAM) microphone is a sensor that is attached behind the speaker's ear [18]. It is able to capture speech which is uttered very quietly (non-audible murmur), through the body tissue. It is used in situations when privacy in communication is preferable.

Approaches to process the signals from the above mentioned sensors for various

speech applications are discussed in the following sections.

2.2.2 Approaches to speech enhancement using multiple sensors

In [13], an approach to enhance the degraded acoustic speech used three alternate sensors, the GEMS, P-mic and EGG, along with an acoustic sensor (NM). The alternate sensors were used according to their capacity to represent specific characteristics of speech in different frequency bands. Speech activity detection was performed using the GEMS sensor to detect the voiced speech components. A P-mic (placed on the forehead) was then used to detect the unvoiced speech. The enhancement was achieved by estimating the magnitude and phase components of the short-time Fourier transform of the acoustic and alternate speech signals. Improvement in the quality and intelligibility of the acoustic speech was reported from informal listenings. However, the use of more than one sensor at the throat region may not be suitable for situations where convenience of the user is of primary importance.

In [15, 19], the bone-conduction microphone was combined with the normal microphone for detection and enhancement of the target speaker's speech from a normal microphone signal in the presence of background speech. Speech detection was based on a histogram of the energy in the bone channel. The enhancement of the noisy speech involved learning the speaker-dependent mapping from the bone sensor signal to the clean (target) speech using the SPICE algorithm [20] based on a piecewise linear representation [7]. Then, the bone sensor signal and the noisy signal were combined to estimate the clean speech using the Wiener filter technique. In [16], the clean NM speech is estimated from the combined noisy speech in the normal and bone

sensors using expectation maximization (EM) algorithm. To avoid the prior training of a speaker-dependent model, an algorithm called direct filtering was proposed in [21]. The clean speech was estimated by learning the transfer function from the close-talking channel to the bone-channel from the given utterance using a maximum likelihood framework.

The bone sensor signal is distorted due to teeth clacking (unconscious contact of upper and lower jaw) and noise leakage [22]. The teeth clack and noise-leakage will distort the mapping function between the air and bone channels, resulting in a negative effect on the enhancement. These distortions were removed by making use of the formant trajectories which were estimated from the Linear Prediction (LP) cepstra using the adaptive Kalman filtering algorithm [23]. The speech was then synthesized using the LP cepstra generated using the estimated formant information. The clean close-talking speech was then estimated from a combined noisy close-talking speech and the synthesized bone sensor speech using the maximum likelihood (ML) framework. An improvement over this method in estimating the clean speech was claimed using a graphical model based approach in [24]. The approach used a two-state speech model and an estimated noise model (based on the speech detection) using the correlation between bone and close-talking channels.

The GEMS sensor was combined with the normal microphone to suppress the background noise from the normal microphone speech signal in [14]. The GEMS signal energy was used to obtain a reliable measure of the voiced speech boundaries. The unvoiced speech segments were estimated using the statistics of the user's language. The noise in the remaining non-speech segments of the acoustic speech was suppressed

using the Wiener filter, the Glottal WINdowing (GWIN) filter, and the Glottal CORrelation (GCOR) filter. A similar approach using the GCOR algorithm was proposed in [25].

A Harmonic Comb Filtering technique, that uses the GEMS signal to detect High Signal Power (HSP) locations in the voiced speech spectrum (without harmonicity assumptions) was used for enhancing the perceptual quality of the degraded NM speech in noisy conditions [26]. This system was used in tandem with a Minimum Mean Square Error (MMSE) estimator for voiced speech, while only the MMSE estimator was used for unvoiced speech in [25, 27]. The non-HSP locations were severely suppressed, while the HSP locations were enhanced with the MMSE estimator. The unvoiced speech frames were mildly suppressed to keep the perceptually important cues in the unvoiced speech intact.

The GEMS and P-mic sensors were used to improve the intelligibility of noisy NM speech in the context of a speech coder in low SNR conditions by combining two algorithms, namely, perceptually motivated Constant-Q (CQ) algorithm and an enhanced GCOR algorithm. The CQ algorithm used a perceptually inspired technique for estimation of speech cues. The enhancement gains were controlled using a psychoacoustic masking model [28]. The enhanced GCOR algorithm extracted the desired speech signal that statistically correlated with the glottal excitation supplied by the GEMS from the noisy mixture.

In the above mentioned approaches, the alternate speech sensors were used to improve the perceptual quality of the noisy NM speech. In this thesis, the perceptual quality of the TM speech is enhanced to obtain a speech which is perceptually more

natural than the TM speech.

2.2.3 Speech recognition using alternate speech sensors

Conventional Automatic Speech Recognition (ASR) systems have a good performance in quiet laboratory environments. The performance drops when used in noisy environments and on large vocabularies by stressed speakers, or when used by dialectal speakers. The error rates are too high for most applications. A typical error performance specification of a reliable ASR system is usually in the range of 1 error in 1000 to as low as 1 in 10,000 depending on the application [29]. To reach this goal, factors of 100 to 1000 improvement in speech recognition accuracy in noisy environments are required.

Noise robustness is one of the key challenges to be solved for the ASR systems to be employed in real environments. Various approaches have been proposed to improve ASR using acoustic signals [26]. The use of acoustic signals (from a normal microphone) alone for improving the efficiency of ASR systems would not yield the desired accuracy. One major reason for this is that the acoustic signals contain insufficient information to accurately represent all the sound units of a language [29]. The effects of circulation, speaker variability and noise on the acoustic signal make it difficult to achieve an ASR system that meets the demands for accuracy, cost and speed. Multisensor information (for example, lip movements in addition to speech) have been used to improve the performance of ASR systems in noisy conditions [30]. In this thesis, the throat microphone signals are used for speech recognition studies. Hence, only the multisensor ASR systems that utilize the information from alternate speech sensors to

augment the NM features are discussed in this section.

In an attempt to improve the recognition accuracy of a word recognizer (TI connected digit database) the use of multiple sensors was reported in [31]. The features from accelerometers placed at the throat and nose were combined with the features from a standard microphone, and were vector quantized using an appropriate codebook. An improvement in recognition results using speaker-dependent isolated-word models was reported.

The feature vectors of the TM speech and the noisy NM speech were combined to estimate the feature vectors of the clean NM speech that are used by standard speech recognizers [32]. A probabilistic optimum filter (piecewise linear transformation) was used to map the temporal sequence of noisy mel cepstral features from the normal microphone and the throat microphone, juxtaposed as an extended feature vector, to the clean mel cepstral features from the standard microphone. This study was later extended to take into account the reduction in SNR in the throat microphone in highly noisy environments [33].

Recognition studies on soft whisper recorded with a throat microphone using various adaptation methods were reported in [34]. Such a system is useful when privacy in human-machine communication is desired. A standard speech recognizer was used. The training data (normal speech) was adapted to the testing domain (whispered speech) for retraining. The adaptation methods included speaker-dependent maximum likelihood linear regression and feature space adaptation, and retraining with downsampling, sigmoidal low-pass filter, and linear multivariate regression. A similar speech recognition study on non-audible murmur in clean and noisy environments

using NAM microphones was reported in [18].

Providing acoustic-phonetic knowledge to speech processing applications would help improve their performance. Towards achieving this, an approach to recognize the broad phoneme classes using the fused features of a standard acoustic signal and a GEMS signal in noisy conditions was reported in [35]. An HMM-based syllable recognition system using the fused features (39 features from NM signal- 13 MFCC with their 1st and 2nd order derivatives, and 3 features from GEMS signal - energy and its 1st and 2nd order derivatives) was reported to perform significantly better than the baseline recognition system that used a front end speech enhancement system. A similar fusion approach was used to increase the performance of digit recognition at both low and high SNR in [36]. Noise robust speech recognition using the bone sensor along with the normal microphone to achieve enhancement of the degraded speech prior to recognition was reported in [15].

2.2.4 Speaker recognition using alternate speech sensors

Speaker authentication is a rich area for exploration of multimodal approaches. Traditionally, visual features have been used to supplement speech recognition as well as speaker recognition [37]. These methods supplement the speech features with the features based on the sequence of lip movements for speaker verification. In [38], the use of EGG, GEMS and P-mic sensors along with the standard microphone was proposed to augment a closed-set speaker recognition system. Speaker specific features were extracted from each of the sensor signals. Gaussian Mixture Models (GMM) and Support Vector Machines (SVM) were then used for the speaker recognition task. A late

integration of the different classification systems was reported to significantly improve speaker recognition in noise.

The GEMS signal, when combined with the acoustic signal from a normal microphone for a speaker verification system, was reported to achieve a 10 fold reduction in error rates under moderate noisy conditions [39]. Apart from the cepstral coefficients derived from the acoustic signal, the GEMS signal was used to derive additional features such as pitch, GEMS shape parameter, and Auto Regression and Moving Average (ARMA) coefficients. This set of feature vectors was used in the dynamic time warping algorithm to calculate the performance "distance" used to make the accept/reject decision on an identity claim.

2.2.5 Speech encoding using alternative speech sensors

The relatively immune signals from the alternate speech sensors supplement the acoustic speech waveform in improving the intelligibility of the speech transmitted through low-rate coders in highly noisy conditions. The signals from the GEMS, P-mic and bone-conduction sensors were fused with the acoustic signals to achieve better intelligibility performance over the standard 2400-bps MELP coder [17]. Pitch and voicing parameters used by MELP were estimated from the GEMS and P-mic signals rather than the acoustic signals. The remaining parameters were obtained from the enhanced (noise suppressed) fused signal.

The GEMS sensor was used in speech compression, where an almost 10-fold bandwidth reduction was reported in comparison to a standard 2.4 kbps LPC 10 protocol [40]. The GEMS signal was used to classify the speech into voiced, unvoiced and si-

lence segments (during which no coding is needed). The excitation function from the GEMS data was used to compute a short-time transfer function for using an ARMA model. As variation in the motion of the vocal articulators is slow, the corresponding rates of variation in the poles and zeros in the complex plane are also slow. This was used to compact the information into a small number of bits/sec. Reduction in transmission bit rate and improved intelligibility in noisy conditions was reported in [41] using a trajectory compression technique, combined with multi-sensory inputs from the standard microphone, EEG and GEMS sensors. The trajectory compression using polynomial approximation exploited the inter-frame information redundancy in natural speech. The EGG and GEMS signal were used to estimate the pitch and voicing parameters as in [17].

The application of alternate speech sensors for various speech systems were discussed in these sections. Most of the reported works dealt with the task of using suitable features from these sensors to improve the robustness of existing speech systems in noisy conditions. In this thesis, the goal is to develop speech systems using the throat microphone speech alone. In order to improve the perceptual quality of the TM speech, the existence of information redundancy in the TM speech and NM speech of a speaker is exploited.

A lot of interest has been evinced in recent years to improve the perceptual quality of the narrowband telephone speech by extending its bandwidth. In this thesis, the approach to improve the perceptual quality of the TM speech is extended for bandwidth extension of the telephone speech. The previous efforts towards bandwidth extension of telephone speech are discussed in the following section.

2.3 APPROACHES TO BANDWIDTH EXTENSION OF NARROWBAND TELEPHONE SPEECH

The basic idea of bandwidth extension algorithms is to extract information about the missing speech components from the available narrowband signal (refer Section. 1.4). Most of the techniques used for this task employ the source-filter model of speech generation. In these techniques, the reconstruction problem is divided into two separate tasks. The first task is recreating a set of wideband Linear Prediction (LP) coefficients, and the second task is forming the wideband residual error signal. The correlation between the highband (4-8 kHz) and the lowband (0-4 kHz) frequency components are exploited. Once the wideband spectral envelope has been generated, the wideband residual is used to excite the wideband LP synthesis filter resulting in a regenerated wideband speech signal. In the ideal case, it is desirable to recover the speech components between 3.3 kHz and 8 kHz, as well as components in the low (0-300 Hz) frequency band [42].

2.3.1 Techniques to regenerate wideband spectral envelopes

2.3.1.1 Linear mapping techniques

The linear mapping between the narrowband spectral vector $\mathbf{x}$ and the wideband spectral vector $\mathbf{y}$ is given by

$$\mathbf{y} = \mathbf{W}\mathbf{x}, \tag{2.1}$$

where the elements of transformation matrix $\mathbf{W}$ are determined using least squares [43]. The matrix $\mathbf{W}$ is calculated as

$$\mathbf{W} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{2.2}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are the matrices comprising the narrowband and wideband spectral vectors respectively.

In [44], Line Spectral Frequencies (LSF) were proposed for estimating the wideband spectral envelope. The linear mapping from the narrowband LSF to the wideband LSF was achieved using multiple matrices. The matrices corresponded to the different clusters of the speech frames based on the spectral shape. In [45], the cepstral coefficients of the narrowband (downsampled wideband) speech were mapped to the cepstral coefficients of the wideband (sampling frequency = 16 kHz) speech by means of a linear predictor that was trained over generic speech files. In [46], the wideband spectral envelope was estimated by mapping the time trajectories of the cepstral coefficients of the narrowband and wideband speech using a filterbank technique.

2.3.1.2 Codebook mapping techniques

The wideband spectral envelope has been estimated using codebook mapping. Codebook mapping relies on a one-to-one mapping between codebooks of narrowband and wideband spectral envelopes to predict the wideband envelope. The wideband spectral envelope was determined from the wideband code vector whose corresponding narrowband code vector is closest in shape to the spectral envelope of the frame of input narrowband speech under analysis in [43] . Euclidean norm was used to find the best match in the narrowband codebook. In [47], the Itakura-Saito and the Gardner-Rao distortion measures were reported to improve the correlation between the distance metric and human perception of difference in the reconstructed speech frames. An optimal gain was applied to the wideband synthesis filter to ensure that the estimated

wideband information had the appropriate energy. Instead of using LP-derived features, MFCC were used for vector quantization in [48].

Using the codebook mapping scheme, the number of possible wideband envelopes which can be predicted is limited to the size of the codebooks. To increase the number of possible estimates of wideband envelopes, two schemes were implemented in [49]. In the first scheme, N narrowband code vectors that were closest to the input narrowband envelope were chosen. The weighted combination of the corresponding N wideband code vectors was used as the estimated wideband envelope. In the second scheme, two separate codebooks for voiced and unvoiced spectral envelopes were used. This split codebook mapping was reported to produce a smaller spectral distortion than the interpolation scheme. In [50], the following techniques were proposed to improve the performance of codebook mapping: (a) Weighting a distance measure towards increased phonetic classification, (b) marginal LSF interpolation, (c) codebook mapping with memory, and (d) codebook interpolation. The weighting for the distance measure depended on the mutual information (1 bit) between the short time critical-band logarithm spectral energy and phonetic classification. The memory information was incorporated into the codebook mapping by the distance between the narrowband LSF of the current and previous frames to interpolate the previous wideband LSF to estimate the current wideband LSF. In [51], one scheme used the distance between the upsampled narrowband LSF and the selected narrowband LSF code vector to estimate the wideband LSF. Another scheme interpolated the upsampled narrowband LSF and the mapped wideband LSF.

2.3.1.3 Statistical mapping techniques

The statistical dependence between the narrowband and wideband speech spectra is exploited to estimate the wideband spectra based on probabilistic measures. In [52], spectral envelopes were modelled as combinations of signals emitted from different random sources. Each source generated autoregressive spectral envelope parameters based on a Gaussian distribution. The narrowband and wideband source were correlated according to a transition probability matrix $P = [p_{ij}]$, where p_{ij} was the probability that the wideband speech was generated by the jth wideband source given that the narrowband portion was generated by the ith narrowband source.

Narrowband spectral envelope of input speech was transformed to wideband spectral envelope based on the Gaussian Mixture Model (GMM) in [53]. The mapping function for this transformation was the least squares regression estimate, obtained from (narrowband and wideband) training data. In [54], a unified probabilistic framework was created which integrated the feature denoising and bandwidth extension process using a single shared statistical model. The GMM trained on narrowband speech and a state-conditional affine transform in the Mel Frequency Cepstral Coefficients (MFCC) domain to transformed the narrowband spectral envelope into a wideband spectral envelope. Hidden Markov Model (HMM) was exploited to indicate the proper representatives of different speech frames to improve the performance of LSF-based approaches by applying a minimum mean square criteria to estimate the wideband LSF values [55, 56]. In [57], equalization was combined with statistical estimation for wideband spectral estimation, and an improvement was reported over the

statistical estimation techniques.

2.3.1.4 Neural networks-based techniques

Neural network techniques to extract the missing frequency contents by a simple nonlinear mapping from narrowband to wideband speech signal [58] used a multilayer perceptron in feedforward operation with three layers for mapping. Adaptive spline neural networks were used for wideband estimation in [59], which was reported to reduce the computational load of standard neural networks, and improve the generalisation capabilities.

2.3.2 Generation of wideband excitation signal

This section briefly explains some of the widely used techniques for generation of the wideband excitation signal. The modulation technique for wideband residual regeneration is to multiply the residual signal with a modulation function. The resulting signal is filtered with a band-stop filter that is designed corresponding to the telephone bandpass. The filtered signal is added to the original narrowband excitation signal [4].

The nonlinear processing technique applies a nonlinear transformation such as a full-wave rectifier or a square operation to the narrowband residual. This transformation creates high frequency components that have a continuous harmonic structure with the baseband. The resulting signal is spectrally flattened so that the excitation does not affect the overall spectral shape [60].

The spectral folding method generates folded images of the baseband spectrum by inserting zeros between each sample of the signal. This upsampling is applied to the LP residual signal so that only the time structures are replicated in the highband

[61]. A pitch synchronous analysis was used in [62] to compute the glottal closure instants. Wideband residual regeneration was based on time-scaling the open phase of the glottal source waveform. This transformed the open quotient, which is a time-domain parameter of the glottal flow signal.

In the above mentioned techniques, the speech synthesized using the wideband residual and estimated wideband spectral envelope was highpass filtered. The original narrowband signal was upsampled by a factor of two and then lowpass filtered. The two signals were then added to obtain the wideband speech.

2.4 OUTLINE OF THE WORK PRESENTED IN THIS THESIS

Previous work in processing speech from alternate speech sensors mainly focused on extracting features from the alternate speech sensors to be used in tandem with the features extracted from the NM speech for improving the performance of the existing speech systems. In the work presented in this thesis, the focus is on processing the speech from a throat microphone for improving its perceptual quality, and exploiting its robustness in noise to develop speech systems based only on the TM speech. Improvement in the perceptual quality of the TM speech would help alleviate discomfort of listeners of the TM speech. Building speech systems based on the TM speech is useful in situations where normal microphones cannot be used. The approach used for improving the perceptual quality of the TM speech is extended for bandwidth extension of telephone speech and loudness enhancement of soft voices.

Chapter 3 analyses the vocal tract characteristics and excitation source characteristics of the TM speech in comparison with that of the NM speech. The approaches to

improve the perceptual quality of the TM speech are detailed in Chapters 4 and 5. The speech applications based on TM signal are discussed in Chapters 6 and 7. Chapter 8 discusses the bandwidth extension of telephone speech and loudness enhancement of soft voices.

2.5 SUMMARY

In this chapter, a review of the approaches used to process speech from alternate speech sensors for various applications like speech enhancement, speech recognition, speaker recognition and speech coding was presented. The approaches to bandwidth extension of the narrowband telephone speech were also reviewed. The outline of the work presented in this thesis was also given.

CHAPTER 3

ACOUSTIC ANALYSIS OF THROAT MICROPHONE SPEECH

3.1 INTRODUCTION

The perceptual differences in speech obtained from the Throat Microphone (TM) and the Normal Microphone (NM) depend on the acoustic characteristics of the sound units in the two speech signals. The acoustic characteristics of the normal microphone speech have been extensively studied in literature [63–76]. This chapter presents studies on the acoustic characteristics of the sound units in the TM signal in comparison with those observed in the NM signal. The study analyses the acoustic waveforms, spectrograms, pitch-synchronous linear prediction (LP) spectra and formant trajectories of syllables [77] to derive inferences. As a result, some of the distinct acoustic features of the TM speech are identified. This identification would be useful to exploit the characteristics of the TM speech for various speech applications.

In the pitch-synchronous analysis, the segments are selected around the instants of significant excitation of the vocal tract system. The regions around the instants are chosen because the significant excitation of the vocal tract takes place during the rapid closing phase of the periodic glottal vibration. The characteristics of the vocal tract system are preserved in the signal in the closed glottis region. Analysis of the speech signal over such an interval provides an accurate estimate of the frequency response of the vocal tract system. A short (2-3 ms) segment to the right of the

instant corresponds to the closed glottis region, and a short (2-3 ms) segment to the left of the instant is considered as the open glottis region. In this chapter, the vocal tract system characteristics of voiced sounds in the closed-glottis region are studied using pitch-synchronous analysis.

3.2 VOWELS

Vowels are produced by exciting a fixed vocal tract shape with quasi-periodic pulses of air flow through the vocal folds, which are vibrating at the fundamental frequency. The cross-sectional area of the vocal tract varies for different vowels. This variation determines the resonant frequencies of the vowels. The tongue is the primary articulator whose position alters the cross-sectional area of the vocal tract, and consequently determines the vowel that is produced. The positions of other articulators such as the jaw and lips also influence the sound produced. Vowels are classified as front, central or back, based on the position of the tongue hump. Vowels are also classified as high, mid and low, based on the height of the tongue hump. The height of the tongue hump is with reference to the roof of the mouth . The vowel is high when the tongue hump is nearest to the roof of the tongue.

The duration of the vowels is longer compared to that of the consonants. The vowels are well defined in the TM and the NM speech. Vowels are distinguished primarily by the location of the first three formant frequencies. Though higher formants exist, they are not necessary for the perception of vowel differences [78]. Hence, only the first three formants are analysed here.

During the production of front vowels */i/* and */e/*, the tongue is pulled forward

towards the front of the oral cavity. This causes a large cavity in the back of the mouth. This large body of air is the source of the low first resonant frequency (F_1), since large objects have low resonances. The cavity in front of the tongue hump is very small. This is the source of the high second resonant frequency (F_2). The low F_1 and high F_2 in front vowels are observed in both NM and TM speech for both */i/* and */e/* (refer Fig. 3.1 (a) and (b)). As the tongue height reduces from */i/* to */e/*, there is a rise in F_1. This rise in F_1 shows that the tongue height is inversely related to F_1, observed in both the speech signals. The decrease in F_2 from */i/* to */e/* is due to the increase in the cross-sectional area of the front cavity from */i/* to */e/*.

When the mid vowel */a/* (as in '*part*') is produced, the tongue hump is low and approximately in the center of the oral cavity to form a large oral cavity. The change in cross-sectional area occurs near the midpoint of the oral cavity [78]. A higher F_1 is observed compared to that of the front vowels, since the size of the cavity behind the tongue hump is smaller when producing */a/* as compared to producing */e/*. The F_2 is very close to F_1 in the NM speech. In the TM speech, F_2 is not as close to F_1 as in the NM speech (refer Fig. 3.1 (c)). The lowered second formant in the NM speech is largely because of the effect of slight lip rounding (compared to the wide spread lips during the utterance of the front vowels). Lip rounding has a lesser effect in the TM speech. Hence F_1 and F_2 are farther apart.

When the back vowels */u/* and */o/* are uttered, the tongue is pulled back. The tongue hump is high for */o/* and lower for */u/*. The most important aspect associated with the production of the back vowel is the rounding of the lips. The position of the jaw affects the back vowels, unlike the front vowels. The lips are protruded when

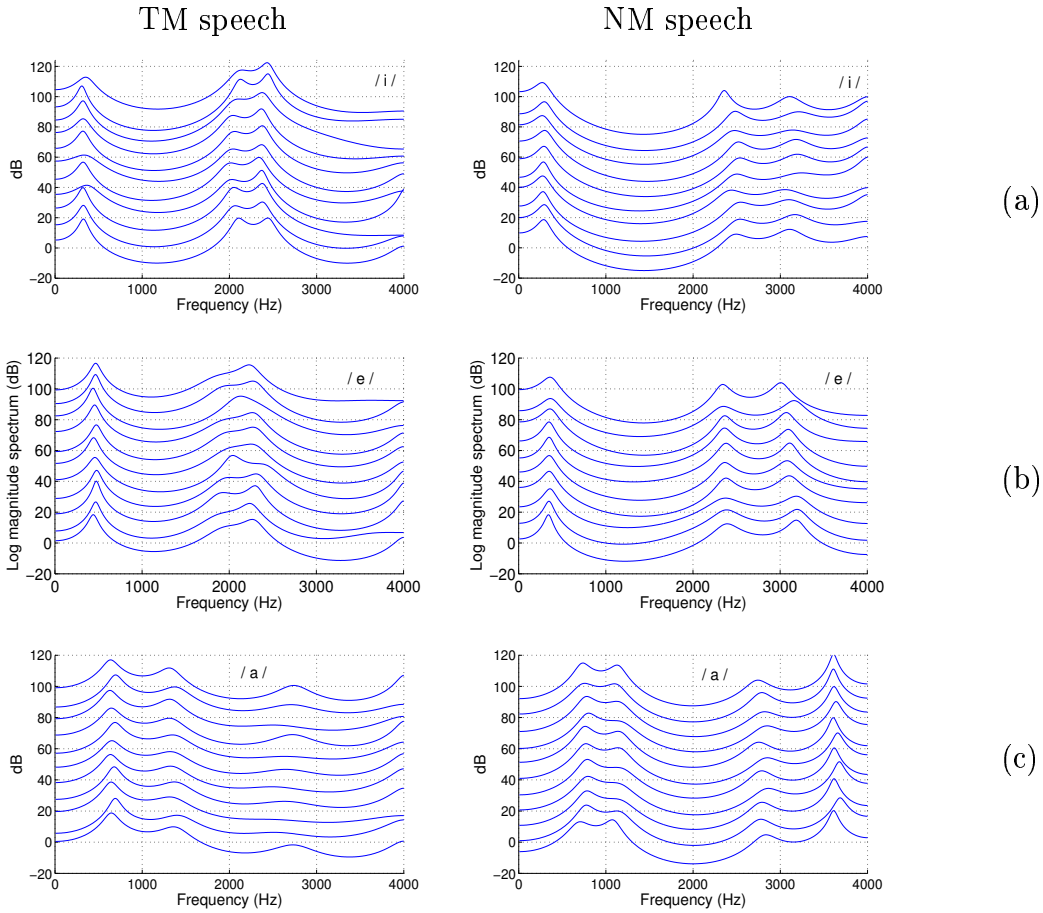

Figure 3.1: LP spectra of 11 successive closed-glottis regions of front vowels (a) */i/*, and (b) */e/*, and (c) mid vowel */a/*, from simultaneously recorded TM speech and NM speech.

rounded. The protrusion increases the length of the oral cavity. This causes the lowering of F_2 in the NM speech (refer Fig. 3.2) [79]. However, as mentioned earlier, the rounding of lips does not affect F_2 in the TM speech. This results in a high F_2 in the TM speech. The locations of F_1 and F_2 are similar in the front and back vowels in the case of TM speech (refer Figs. 3.1 and 3.2).

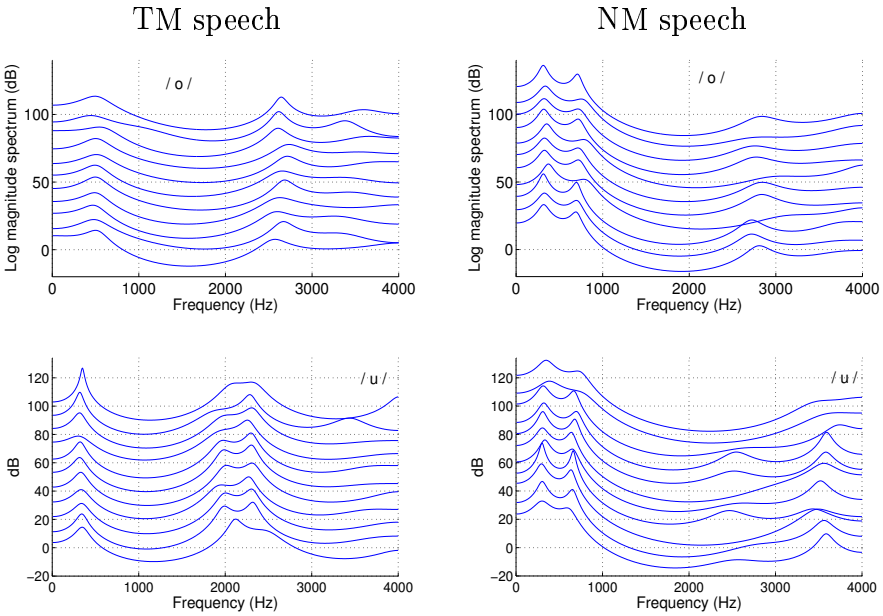

Figure 3.2: LP spectra of 11 successive closed-glottis regions of back vowels (a) */o/*, and (b) */u/*, from simultaneously recorded TM speech and NM speech.

3.3 STOP CONSONANTS

Stop consonants can be categorised into two classes; unvoiced and voiced. In each class there are two categories; unaspirated and aspirated. With the stop consonants, the sequence of events is as follows [63]:

1. There is a complete closure of the oral tract at some location (hence the name 'stop consonants') and build-up of air pressure behind the closure. During this time of closure, no sound is radiated from the lips. For voiced stops, along with the build-up of air pressure, the vocal folds also vibrate.

2. Release of air pressure and generation of turbulence over a very short time duration, i.e., the burst source, which excites the oral cavity in front of the constriction.

3. Generation of aspiration (in the case of aspirated stops) due to turbulence at the open vocal folds (before onset of vibration) as air rushes through the open oral cavity after the burst.

4. Onset of the following vowel after the burst.

During the closure phase, unvoiced stops (e.g., */k/*) are characterised acoustically by the absence of energy in the NM signal as well as the TM signal, as seen in Fig. 3.3.

During the closure phase of voiced stop consonants (e.g., */g/*) in the NM speech, there is no radiation of sound from the lips. The vocal fold vibrations during the closure are propagated through the walls of the throat, which are captured by the normal microphone. These low frequency vibrations are seen as an energy band in the 0 to 500 Hz range. This is called the "voice bar", and is similar for all the voiced stops [79]. However, each of the stop sounds has its effect on the adjacent vowel. At the onset of the adjacent vowel, the formants present correspond to the particular shape of the vocal tract after the release of the closure. Figure 3.4 shows the spectrograms of the syllables, */ga/, /ja/, /Da/, /da/* and */ba/*. In all the voiced stops, the first formant rises from a low position. This is a mark of a stop closure and does not play a role in distinguishing one place of articulation from other.

In */ga/*, F_2 is steady, while F_3 rises from a low locus. For */ja/*, both F_2 and F_3 decrease from the high locus. In */Da/* and */da/* F_2 decreases, with the locus for */Da/* being higher than in */da/*. The F_3 appears to be steady in both the stops. The onset of F_2 and F_3 is lower in */ba/* than in */Da/*. The F_2 is low in */ba/* due to the lip rounding

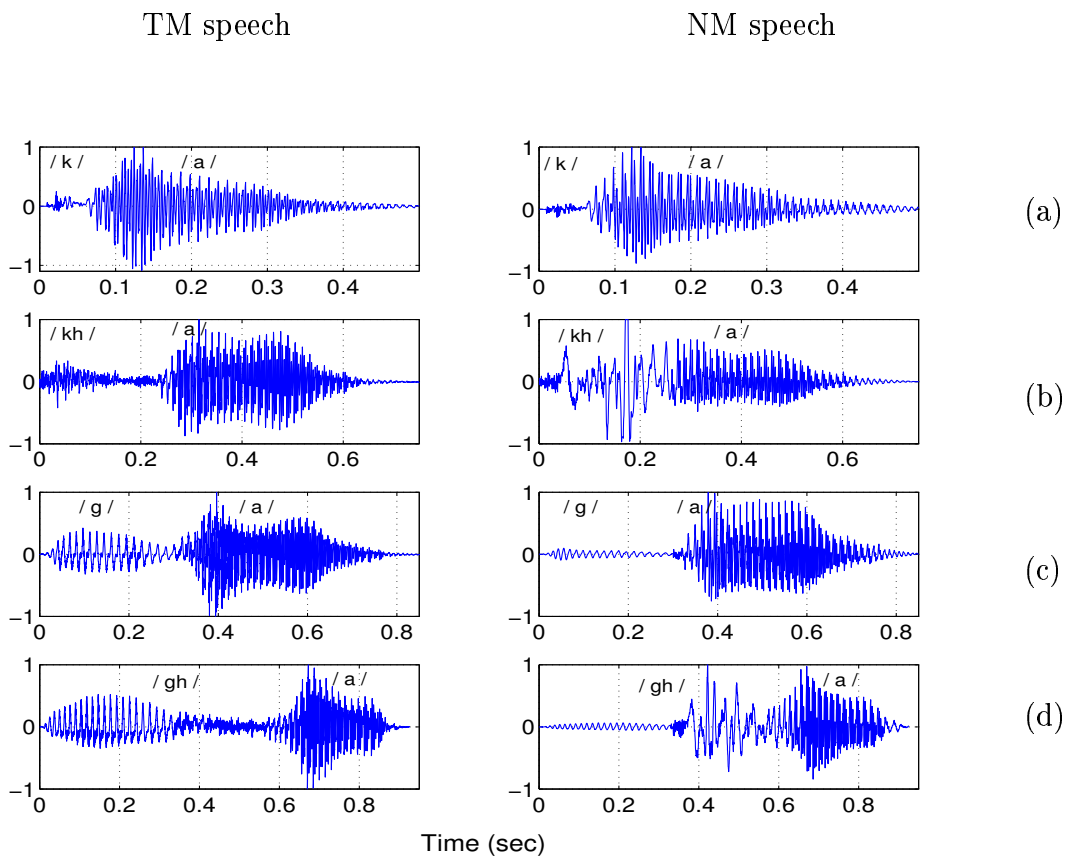

Figure 3.3: The speech signal waveform for the syllables (velar stop consonant-vowel units) (a) */ka/*, (b) */kha/*, (c) */ga/* and (d) */gha/* recorded simultaneously using a throat microphone and a normal microphone.

during closure. It rises from a low locus and moves towards the steady region. There is only a small rise in F_3.

In contrast to the voiced stops in the NM speech where the distinguishing features are the loci of the formants of the following vowel, the voiced stops in the TM speech are distinguished by the presence of formant-like structures in the closure phase of the voiced stops. The vibrations of the vocal folds during the closure result in resonances of the oral cavity behind the place of closure. As the TM is placed near the pharynx

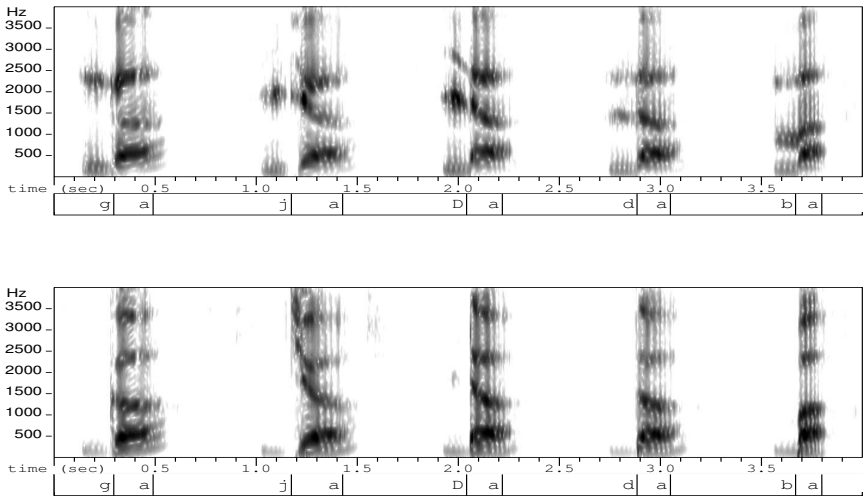

Figure 3.4: The spectrograms for the syllables */ga/, /ja/, /Da/, /da/, /ba/* recorded simultaneously from a throat microphone (above) and a normal microphone (below).

region, it is able to capture the resonances associated with the closure, unlike the NM. A low frequency band like the voice bar is seen in all the voice stops, similar to NM speech, indicating the mark of a stop closure. In addition, the consonants */g/* and */j/* have a second formant, while the consonants */D/, /d/* and */b/* have a third formant as well. The F_2 progressively decreases as we move from */g/* to */b/*. The F_2 ends where the onset of the F_2 of the adjacent vowel begins. The F_3 in */D/* decreases towards the onset of the F_3 of the adjacent vowel. The F_3 appears to be steady in both */d/* and */b/*, with the F_3 in */d/* being at a higher location than in */b/*. These distinct formant-like structures associated with closure of the voiced stops could serve as acoustic cues to resolve voiced stops into places of articulation classes. The pitch-synchronous formant trajectories for the syllables */agag/, /aDaD/* and */abab/*

are shown in Fig. 3.5. It is seen that the formant frequencies are present for the entire syllable in the TM speech, while the formant frequencies are present only in the vowel regions in the NM speech. This clearly shows that the resonances of the vocal tract during the closure are captured by the TM.

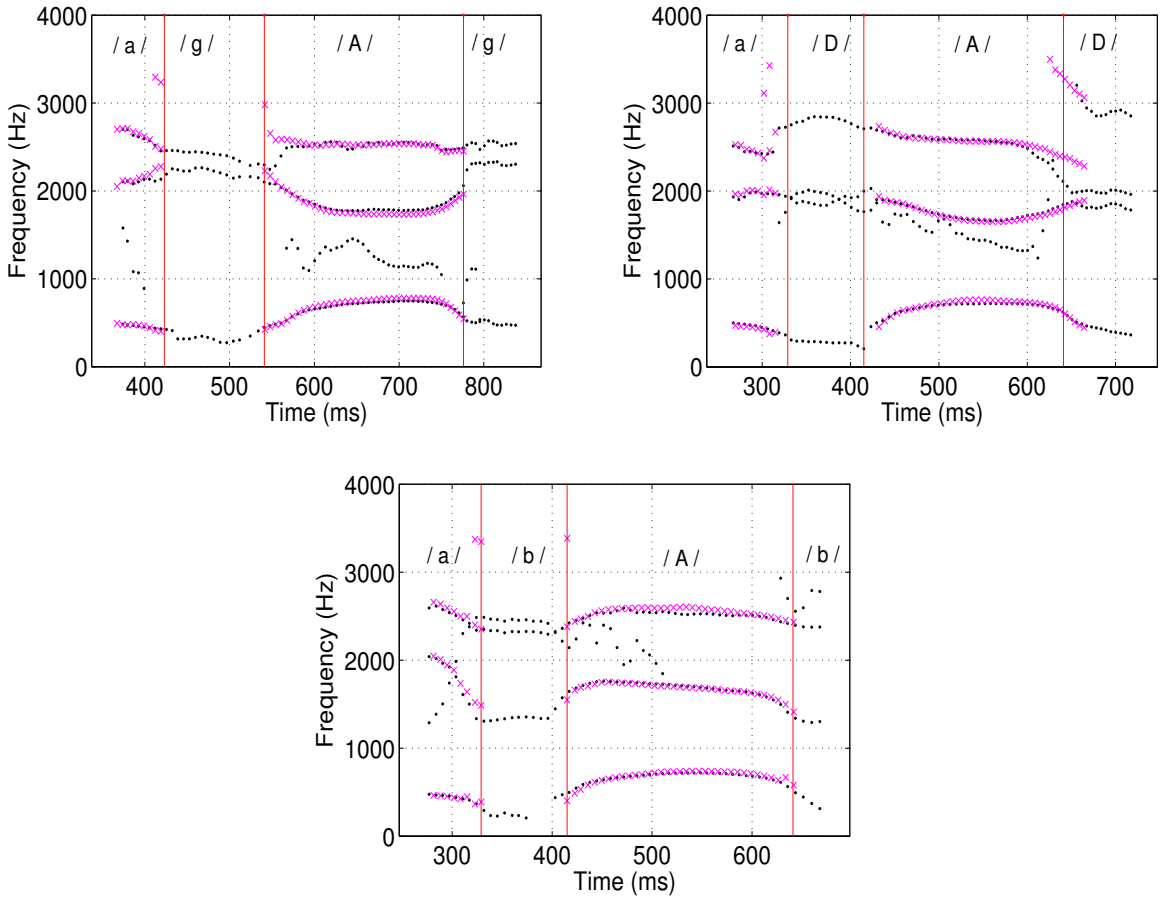

Figure 3.5: The formant trajectories for the VCVC phrases */a gag/*, */a DaD/* and */a bab/*, recorded simultaneously using a throat microphone (shown as '•') and a normal microphone (shown as '×').

3.4 NASAL CONSONANTS

Nasal consonants are produced with the glottal excitation, and the velum lowered to allow the coupling of the nasal and oral tracts. There is complete closure of the oral

tract at a location dependent on the nasal sound. The nasal consonants are radiated through the nostrils. The oral tract acts as a side branching resonator that selectively absorbs energy from the main tube at frequencies that are dependent on the oral resonances [63]. The waveforms of nasals in the NM speech are characterized by low amplitude, periodic, highly damped structure. However, for the TM speech, the nasal waveforms have a structure similar to that of the vowels. The damping is very less since the damping effect of the nasal cavity is not captured well by the TM. This can be seen in the waveform segments of the vowel-consonant-vowel syllables */ama/* and */ana/* in Fig. 3.6.

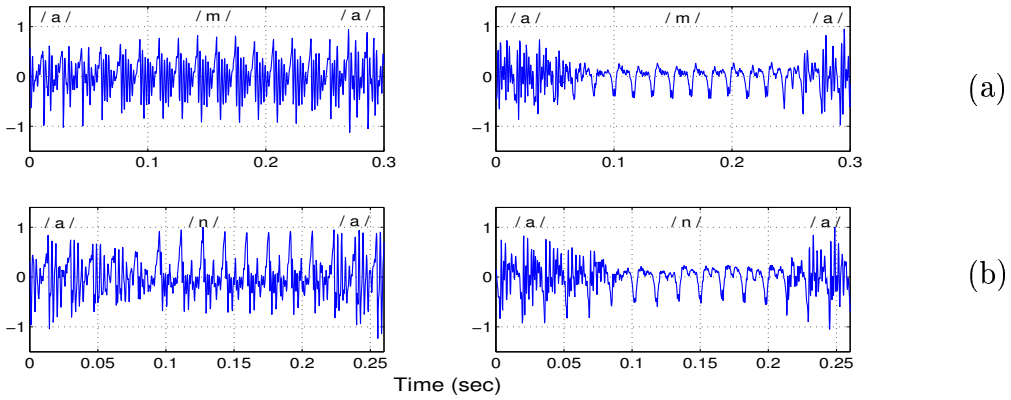

Figure 3.6: The waveforms for the syllables */ama/* and */ana/* recorded simultaneously using (left) a throat microphone, and (right) a normal microphone.

The acoustic structure of the nasal consonants is dominated primarily by the resonances of the nasal-pharyngeal tube and the anti-resonances of the mouth cavity. The nasals have a formant structure similar to that of a vowel, but are in particular frequency locations that depend on the characteristic resonances of the nasal cavities. Also, the nasal consonants are low in energy when compared to the vowels.

In the NM speech the nasal consonants have a very low first formant centered at about 250 Hz (as seen in Fig. 3.7). The locations of the higher formants vary, but generally there is a large region above the first formant with no energy. In the TM speech, due to the location of the sensor, the oral resonances are also seen in the spectrum of the nasal consonants. The oral resonance for */n/* is at 1400 Hz, and for */m/* is at 900 Hz. The oral resonance in both */n/* and */m/* shows some continuity with the second formant of the adjacent vowel.

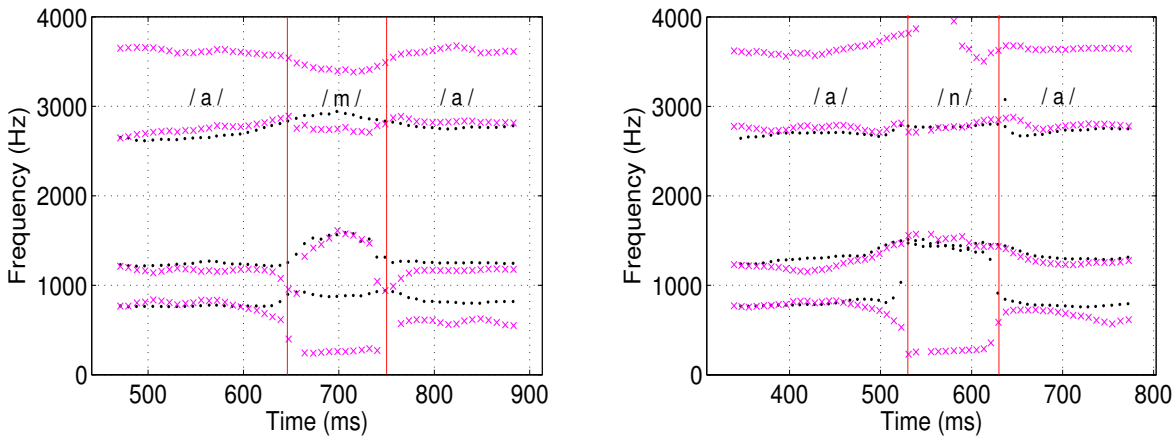

Figure 3.7: The formant trajectories for the VCV syllables */ama/, /ana/* recorded simultaneously using a throat microphone (shown as '•') and a normal microphone (shown as '×').

3.5 FRICATIVES

Fricatives are produced due to turbulence in the stream of air as it passes through a narrow constriction in the oral tract. This turbulent air flow acts as a noise source which excites the cavity in front of the narrow constriction. Fricatives are characterised by the distribution of random energy over a wide range of frequencies. In the NM speech, the highest frequencies (extending beyond 8 kHz) in speech occur during the

production of fricatives. Both */s/* and */sh/* have a large acoustic intensity and hence produce dark patterns. In the TM speech, the fricatives are characterised by the random energy being restricted to a narrow frequency band between 1000 Hz and 3000 Hz (refer Fig. 3.8).

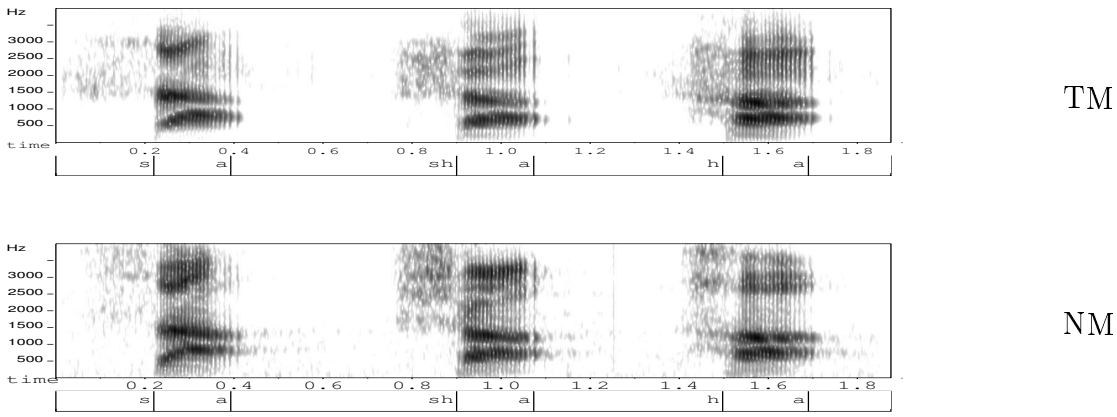

Figure 3.8: The spectrograms for the syllables */sa/, /sha/, /ha/* recorded simultaneously from a throat microphone (above) and a normal microphone (below).

3.6 SEMIVOWELS

Acoustically, semivowels bear the closest similarity to vowels and dipthongs, and are characterised by a formant pattern. The difference in the characteristics of the semivowels in the TM and NM signals is that in the NM signal the acoustic intensity of the semivowels is distinctively lower than that of the vowels, with an abrupt change in intensity observed at the transition from semivowel to vowel (or vice versa). However, in the TM signal, the semivowels have a high acoustic intensity (refer Fig.

3.9). In the NM speech, the first formant of */l/* is around 250 Hz. It rises from */l/* to the vowel target. The second formant is around 1500 Hz, while the third formant is around 2600 Hz. In the TM speech the first formant is around 700 Hz. This could be due to the resonance of the mouth cavity behind the region of (slight) constriction in the oral tract. For the semivowel */y/* the first and second formants are rising and falling respectively in both the TM and NM speech. The difference is observed only in the third formant. It is rising in the TM speech, instead of falling as in the NM speech.

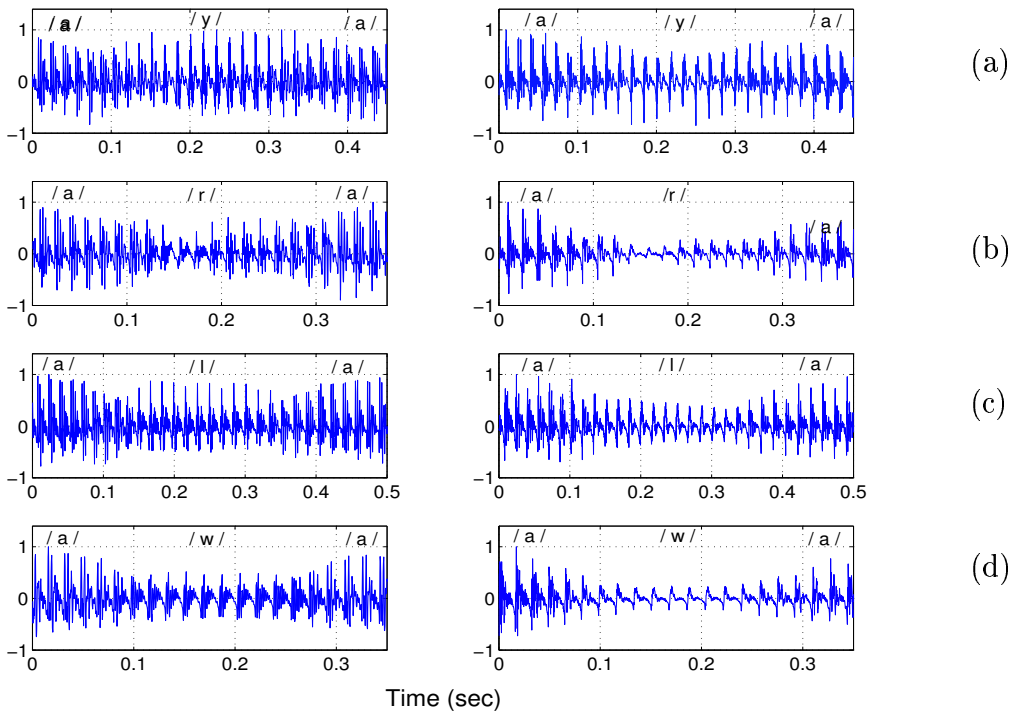

Figure 3.9: The waveforms for the syllables */aya/*, */ara/*, */ala/* and */awa/*, recorded simultaneously using a throat microphone and a normal microphone.

Fig. 3.10 shows the formant trajectories of */aya/* and */ala/*.

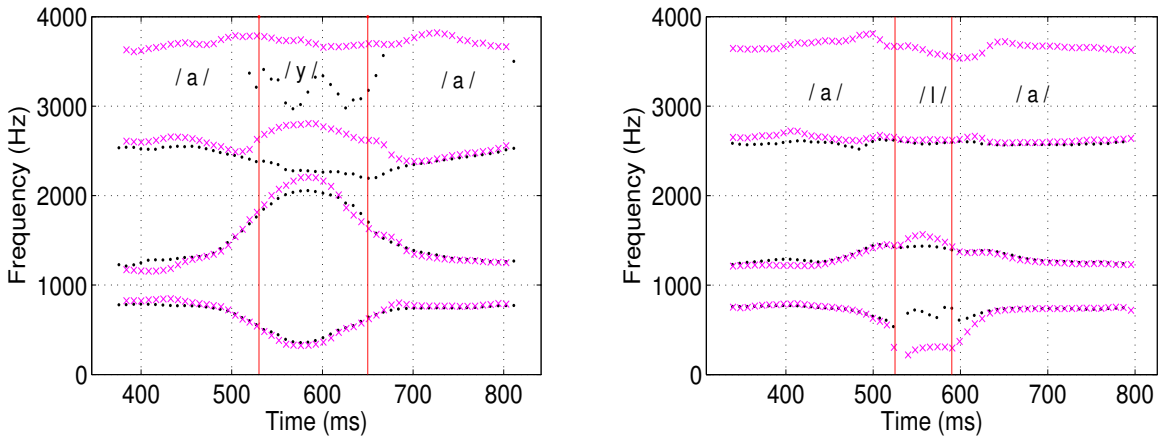

Figure 3.10: The formant trajectories for the VCV syllables */aya/* and */ala/* recorded simultaneously using a throat microphone (shown as '•') and a normal microphone (shown as '×').

3.7 EXCITATION SOURCE CHARACTERISTICS OF VOICED SOUNDS

The excitation source, represented by the LP residual, is characterised by the presence of quasi-periodic impulse-like excitation for voiced sounds. The quasi-periodic impulses correspond to the instants of significant excitation, caused by the closure of the vocal folds. The strength of the instants, which indicates the strength of excitation, mainly depends on the loading of the vocal tract system on the source. In the NM speech, the loading effect is less in the case of vowels due to the unobstructed passage of airflow through the oral tract. Hence, the strength of excitation is significant at the instants compared to the open glottis regions, as seen in Fig. 3.11 (c) and (d). In comparison, the amount of loading is more for voiced consonants like nasals and voiced stops. This is because the articulation of these sounds involves the closure of the oral tract at some location. In the case of nasals, the strength of excitation at the instants is comparable to the strength of the signal in the open glottis regions, as seen in Figs. 3.12 (c) and

3.12 (d). In the case of the voiced stop consonants it is difficult even to define the instants of glottal closure (refer Figs. 3.13 (c) and 3.13 (d)).

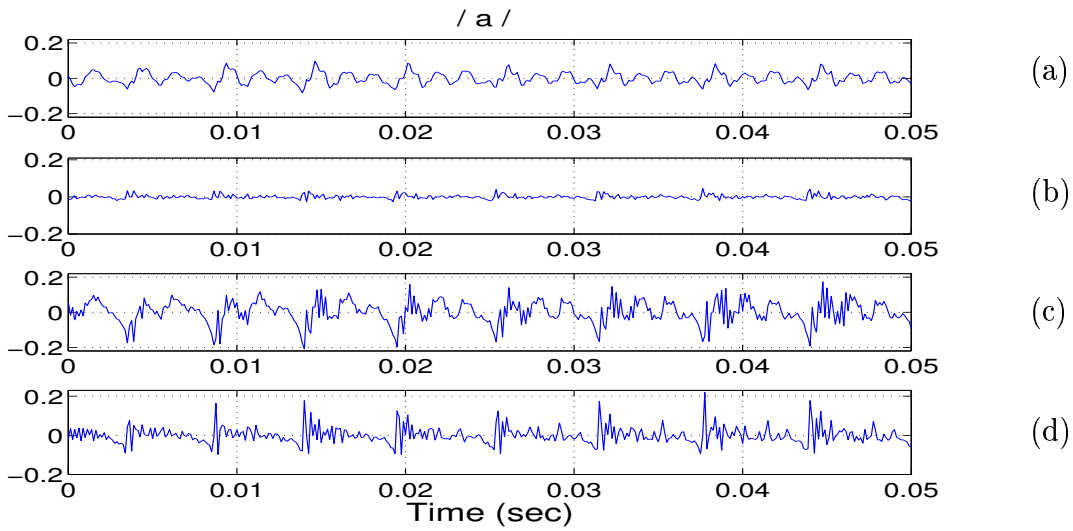

Figure 3.11: (a) TM speech signal waveform for vowel /*a*/, (b) LP residual for the signal in (a), (c) NM speech signal waveform for vowel /*a*/ and, (d) LP residual for the signal in (c).

In the TM speech, the strength of instants for the vowels is less compared to their strength in the NM speech, as seen in Fig. 3.11 (a) and (b). The reduced strength of the instants is because the TM picks up the vowel sounds that are transmitted through the walls of the throat, unlike the unobstructed airflow through the mouth. In the case of nasals, the strength of the instants is higher compared to the strength in the NM speech (refer Fig. 3.12 (a) and (b)). In the voiced stops, low amplitude instants are seen in the TM speech (refer Fig. 3.13 (a) and (b)), compared to the near absence of instants in the NM speech. In the NM speech, the vowel is clearly distinguishable from the voiced consonants based on the strength of the instants. However, in the TM speech, the strengths of the instants are comparable in the vowel and voiced consonant

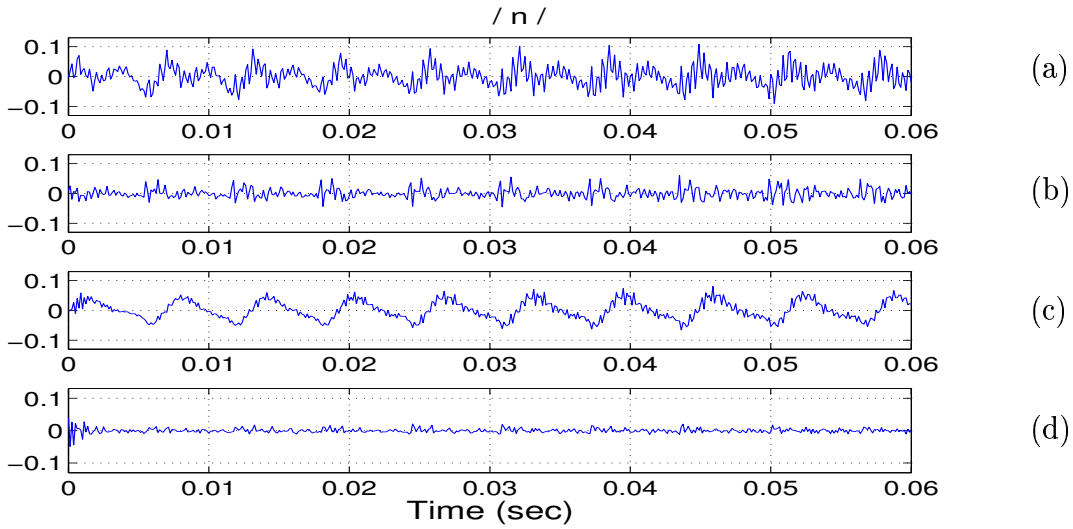

Figure 3.12: (a) TM speech signal waveform for nasal */n/*, (b) LP residual for the signal in (a), (c) NM speech signal waveform for nasal */n/* and, (d) LP residual for the signal in (c).

(nasal) regions.

Fig. 3.14 shows the relative strengths of the instants in the vowel and voiced consonants in the LP residual signals of the TM and NM speech.

Some of the differences in the acoustic characteristics between the TM speech and NM speech for various sound units are summarized in Table 3.1.

3.8 SUMMARY

In this chapter, the acoustic characteristics of various sound units of the TM speech were analysed. This study showed that some higher frequency information is missing, or is of low intensity in the throat microphone signal. This was seen in fricatives and vowels. The pitch-synchronous formant trajectories brought out the differences in the acoustic characteristics of the voiced sound units in the TM and NM speech.

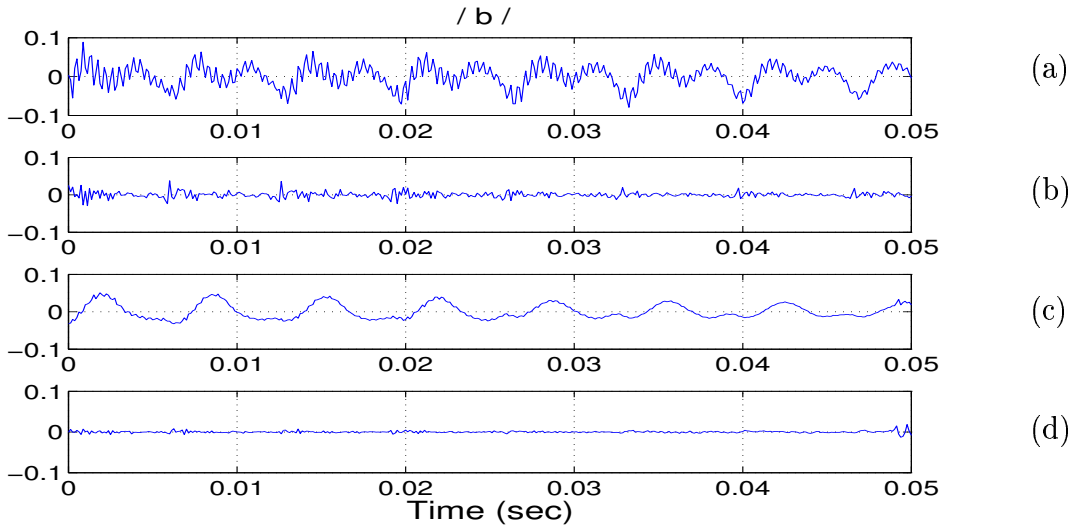

Figure 3.13: (a) TM speech signal waveform for voiced stop */b/*, (b) LP residual for the signal in (a), (c) NM speech signal waveform for voiced stop */b/* and, (d) LP residual for the signal in (c).

This study showed that a discrimination of voiced stop consonants is possible from the acoustic cues present during the articulation of stops in the TM speech. The study also showed that the excitation source characteristics of the TM and NM speech differ in the strength of excitation of the voiced sounds. This study helps to understand the characteristics of the TM signal, and exploit the TM speech for various speech applications, even under adverse conditions.

In the following two chapters, the methods for enhancement of the TM speech are explained.

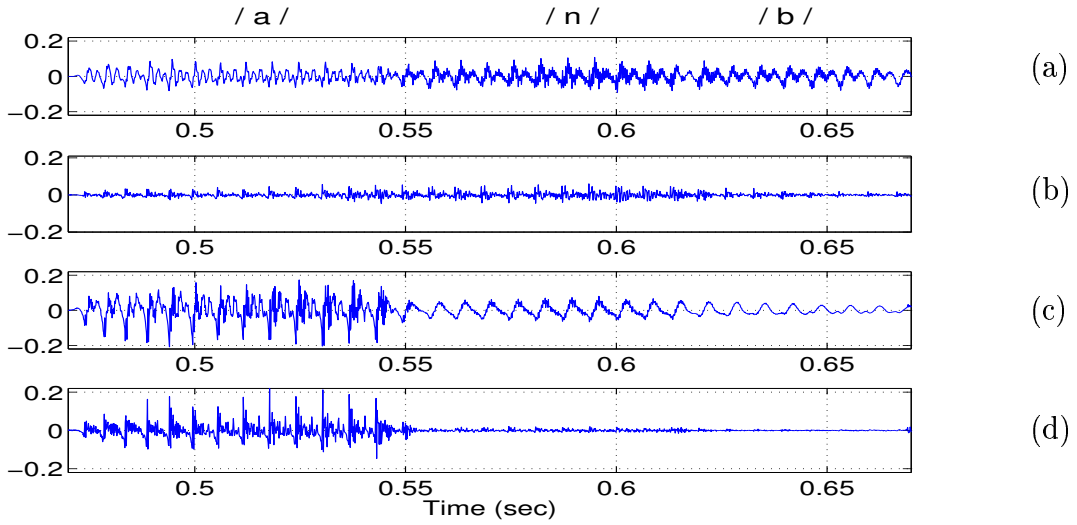

Figure 3.14: (a) TM speech signal waveform for the syllable */anb/*, (b) LP residual for the signal in (a), (c) NM speech signal waveform for the syllable */anb/* and, (d) LP residual for the signal in (c).

Table 3.1: Differences in the characteristics of sound units in NM and TM speech.

Characteristics of sound units	NM Speech	TM Speech
Energy distribution	Entire audio frequency range.	≈ up to 3500 Hz
Formant location of back vowels	Low 2nd formant	High 2nd formant like front vowels
Closure phase of voiced stop consonants	Low frequency "voice-bar"	Formant-like structures
Aspiration phase of stop consonants	Large amplitude noise	Low amplitude noise
Signal damping in nasal consonants	Highly damped	Less damped like vowels
Intensity of formants in semivowels and nasal consonants	Less compared to vowels	Similar to vowels
Formant locations of nasal consonants	Depend on nasal resonances	Higher formant locations depend on oral resonances also
Strength of instants of excitation of voiced segments	High for vowels, low for voiced consonants	Comparable for all voiced sounds

CHAPTER 4

ENHANCEMENT OF THROAT MICROPHONE SPEECH-SPECTRAL MAPPING

4.1 INTRODUCTION

Enhancement of throat microphone speech aims at improving the perceptual quality of the unnatural speech from a Throat Microphone (TM). Though the TM speech is intelligible even in a noisy ambience due to the placement of the sensor near the larynx, the speech sounds unnatural unlike the speech from a normal microphone. Hence the approach for enhancement attempts to improve the naturalness of the TM speech, without affecting its intelligibility. The naturalness of the NM speech is exploited for such an enhancement.

The acoustic analysis of the TM speech discussed in Chapter 3 shows that differences exist between the characteristics of the TM and NM speech signals for various sound units. These differences exist both in the vocal tract characteristics as well as the excitation source characteristics of the speech. The enhancement of the TM speech would, therefore, involve two tasks: (1) Estimating the vocal tract (spectral) characteristics of the NM speech given the TM speech spectra, and (2) estimating the excitation source characteristics of the NM speech, from that of the TM speech. The estimated excitation signal and spectral features are used to reconstruct the enhanced speech. The task of estimating the spectral characteristics of the NM speech is discussed in this chapter, while the task of estimating the excitation source characteristics

of the NM speech is discussed in the next chapter.

This chapter is organized as follows. The proposed method for enhancing the TM speech is discussed in Section 4.2. The issue of stability of the synthesis filter, and the mapping network used for estimation of the NM spectra are discussed in Section 4.3. Section 4.4 discusses the behaviour of the network for different types of sound units, and illustrates the performance of mapping during testing. The objective measure used to assess the quality of the estimated spectra is also discussed in this section. The work is summarized in Section 4.5.

4.2 PROPOSED APPROACH TO SPECTRAL MAPPING

Typically, the throat microphone is used in situations where normal microphones cannot be used (refer Section 1.3.2). In such situations where only the TM speech is available, the spectral features of the NM speech have to be estimated, given the spectral features of the TM speech. Though differences exist in the acoustic characteristics of the TM and NM speech for various sound units, there also exists correlation between the corresponding time frames of the simultaneously recorded TM and NM speech from a speaker. This correlation is speaker-specific because the information about the vocal cord activity (e.g., pitch) and the dimensions of the vocal tract (e.g., location of formants) of a speaker remain similar in both the TM and NM speech. This speaker-specific correlation can be exploited to capture the relationship between the spectral features of the TM and NM speech of a speaker using a mapping technique. Since the information such as pitch and formant locations vary across speakers, a speaker-independent mapping would result in distortions in the synthesized speech.

The spectral features of the TM speech are mapped onto the corresponding spectral features of the NM speech using a neural network-based mapping technique. The mapping involves the following stages (refer Fig. 4.1): The first stage involves training a model to learn the mapping. For the training to be effective, speech is simultaneously recorded using a throat microphone and a normal microphone from a speaker. Simultaneous recording ensures that the model learns the mapping between the corresponding frames of the TM and NM speech. The LP analysis is performed on the speech signals to extract the LP coefficients, from which the weighted LP Cepstral Coefficients (wLPCCs) are derived. For training, the wLPCCs extracted from the TM speech are mapped onto the wLPCCs extracted from the corresponding NM speech. That is, the wLPCCs derived from the TM data are used as input to the mapping network, while the wLPCCs obtained from the NM speech form the desired output. The mapping property of MultiLayered FeedForward Neural Network (MLFFNN) is used to learn this mapping, as described later in Section 4.3.2. The second stage consists of testing, where the wLPCCs derived from a test TM utterance are given as input to the trained network. The output produced by the network are the estimated wLPCCs of the NM speech corresponding to the test input. The LP coefficients are derived from these estimated wLPCCs in order to synthesize the speech.

4.3 MAPPING SPECTRAL FEATURES OF TM AND NM SPEECH

The issues that are addressed in the proposed approach are: (a) Achieving an effective mapping between the spectral features of the TM and NM speech, (b) ensuring that the all-pole filter derived from the learnt mapping is stable, and (c) ensuring that

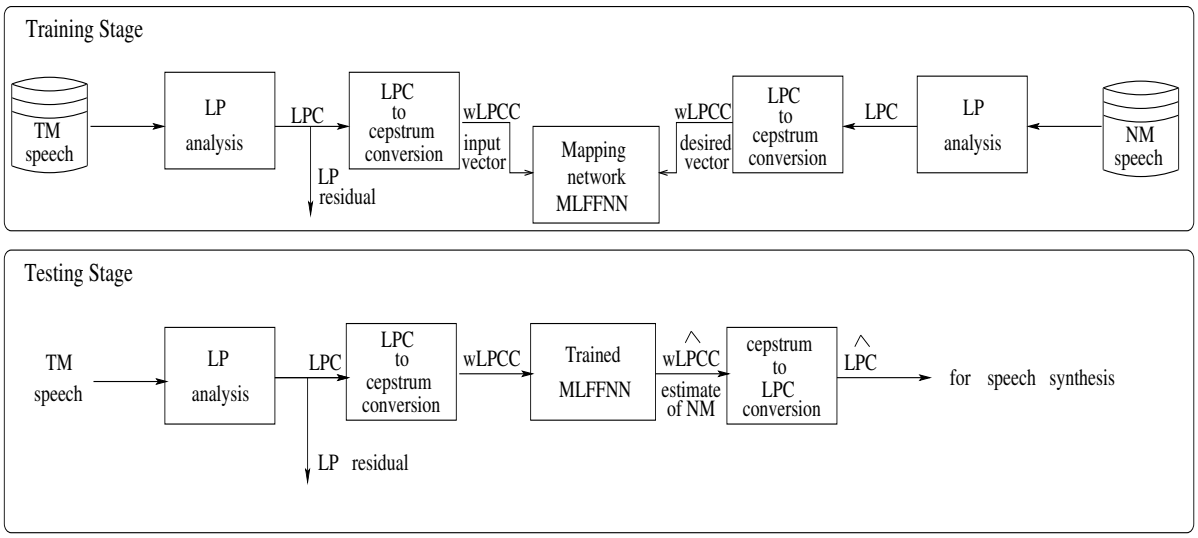

Figure 4.1: Block diagram of the proposed model for capturing the relationship between the spectra of the TM speech and the NM speech of a speaker.

the synthesized speech does not suffer from discontinuities due to spectral 'jumps' between adjacent frames. The filter for synthesis is obtained by (1) using the cepstral coefficients from both the TM and NM speech signals for training a mapping network, and (2) deriving an all-pole filter from the estimated cepstral coefficients from the trained mapping network. The stability of the all-pole filter, discussed in the following section, is essential for the synthesis of the enhanced speech.

4.3.1 Deriving features for mapping and synthesis

Cepstral coefficients are used to represent the feature vector of each frame of speech data. The cepstral coefficients are derived from the LP coefficients as follows:

The LP spectrum for a frame of speech signal is given by

$$|H(k)|^2 = \left| \frac{1}{1 + \sum_{n=1}^{p} a_n e^{-j\frac{2\pi}{M}nk}} \right|^2, \quad k = 0, 1, 2, \ldots, M\text{-}1, \tag{4.1}$$

where $\{a_n\}$ is the set of LP coefficients, and M is the number of spectral values. The

inverse Discrete Fourier Transform (DFT) of the log LP spectrum gives the set of cepstral coefficients $\{c_n\}$. Let

$$S(k) \quad = \quad \log|H(k)|^2. \tag{4.2}$$

Then

$$c_n \quad = \quad \frac{1}{M}\sum_{k=0}^{M-1} S(k)e^{j\frac{2\pi}{M}kn}, \quad n = 0, 1, 2, \ldots, M\text{-}1. \tag{4.3}$$

Only the first q cepstral coefficients are chosen to represent the LP spectrum. Normally, q is chosen to be much larger than the LP order p in order to represent the LP spectrum adequately.

Linearly weighted cepstral coefficients, nc_n, $n = 1, 2, \ldots, q$, are chosen features to represent a frame of the speech signal. The weighted linear prediction cepstral coefficients (wLPCC) are derived for each frame of TM speech and for the corresponding frame of the NM speech. These pairs of wLPCC vectors are used as the input-output pairs to train a neural network model to capture the implicit mapping.

The output of the trained network for each frame of TM data of a test utterance gives an estimate of the wLPCC of the corresponding frame of NM speech. From these estimated wLPCCs, $\hat{c}_n$, $n = 1, 2, \ldots, q$, the estimated log LP spectrum is obtained by using DFT. Let $\hat{S}(k)$, $k = 0, 1, 2, \ldots, M$-1 be the estimated log spectrum. The estimated spectrum $\hat{P}(k)$ is obtained as

$$\hat{P}(k) \quad = \quad e^{\hat{S}(k)}, \quad k = 0, 1, 2, \ldots, M\text{-}1. \tag{4.4}$$

From the spectrum $\hat{P}(k)$, the autocorrelation function $\hat{R}(n)$ is obtained using inverse DFT of $\hat{P}(k)$. The first $p+1$ values of $\hat{R}(n)$ are used in the Levinson-Durbin algorithm

to derive the LP coefficients [80]. These LP coefficients for each frame are used to obtain the time-varying synthesis filter. The enhanced speech is obtained by exciting this time varying filter with the modified LP residual (discussed in the next chapter) of the TM speech.

The all-pole synthesis filter derived from these LP coefficients is stable, and is explained as follows. The stability of the all-pole filter (of the form $\frac{1}{A(z)}$) is determined entirely by the roots of the denominator polynomial $A(z)$ when the filter coefficients are constant [81]. If any root lies outside the unit circle ($|z| = 1$), the filter is said to be unstable, in the sense that it will have an exponentially increasing oscillation for a unit sample input. If a particular analysis frame results in an inverse filter $A(z)$ with roots lying outside the unit circle, a perceptually noticeable effect may be obtained in the synthesized speech, even if other analysis frames have roots within the circle [81].

The LP coefficients are recursively related to the LPCCs by [80]

$$a_n = c_n - \sum_{k=1}^{M-1} \left(\frac{k}{n}\right) c_k a_{n-k}, \quad 1 \leq n \leq p. \tag{4.5}$$

However, the LP coefficients derived from this recursive relation cannot be used for synthesizing speech as the coefficients do not guarantee stability of the filter. This is because the q wLPCCs (typically, truncated to 1.5 times p for computational purposes) used for mapping, though sufficient to represent the LP spectrum, are not sufficient to reconstruct the LP spectrum. The reconstruction of the LP spectrum requires q to be very large (ideally infinite).

The autocorrelation method, used in this work, to derive LP coefficients from the

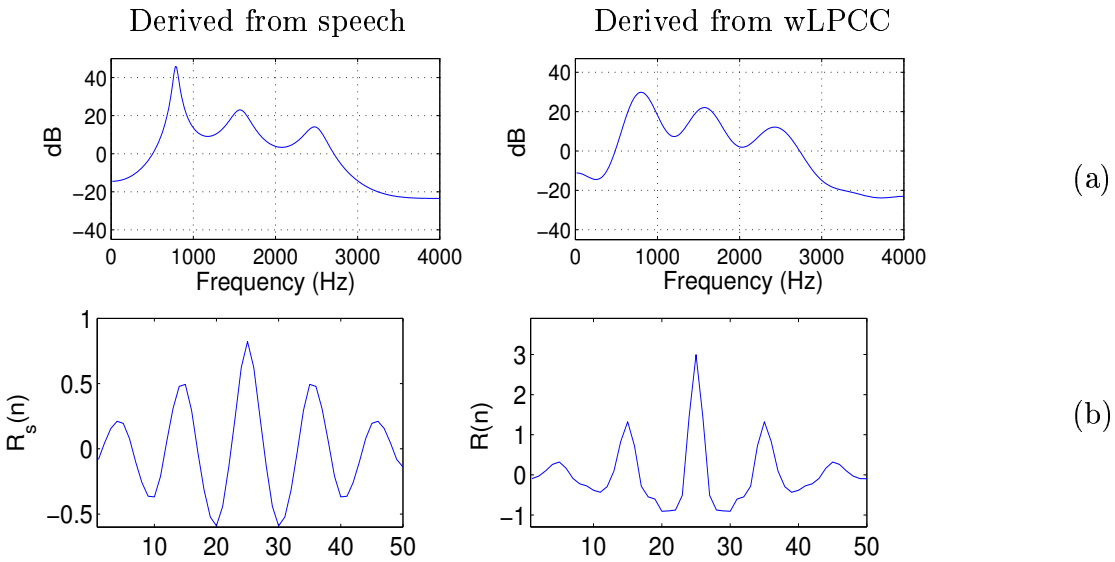

Figure 4.2: (a) The spectrum and (b) the autocorrelation function of a frame of the NM speech derived from (left) the speech signal, and (right) from the wLPCCs of the corresponding frame.

estimated wLPCCs guarantees stability of the filter. This is because the coefficients $\hat{R}_i$ are positive definite [82]. The solution of the autocorrelation equation, given by $\sum_{n=1}^{p} a_n \hat{R}(i-n) = -\hat{R}(i), 1 \leq i \leq p$, would then give predictor parameters which guarantee that all the poles of the filter lie inside the unit circle. The ability of linear prediction to model the speech spectrum is explained by the exact match of the autocorrelation function of the all-pole filter and the autocorrelation of the input signal between the indices 0 and p [83]. As seen in Fig. 4.2, such a similarity is observed between the autocorrelation function ($R(n)$ in figure) derived from the wLPCCs and the autocorrelation ($R_s(n)$ in figure) of the NM speech. The mapping approach is discussed in detail in the following section.

4.3.2 Neural network model for mapping spectral features

The task of capturing the functional relationship between the spectra of the TM speech and NM speech of a speaker is a pattern mapping problem. A neural network approach

is proposed to address the mapping problem. In the pattern mapping problem, given a set of input-output pattern pairs, the objective is to capture the implied mapping between the input and output vectors. Once the system behavior is captured by the neural network, the network would produce a possible output pattern for a new input pattern not used in the training set. The possible output pattern would be an interpolated version of the output patterns corresponding to the input training patterns which is closest to the given test input pattern [84, 85]. The network is said to *generalize* well when the input-output mapping computed by the network is (nearly) correct for the test data that is different from (but close to) the examples used to train the network [86].

Let f denote the spectral mapping learnt by the network.

If $\{(\mathbf{a}_1, \mathbf{b}_1), (\mathbf{a}_2, \mathbf{b}_2), \ldots, (\mathbf{a}_L, \mathbf{b}_L)\}$ is the set of input-output training pattern pairs for which $\{\mathbf{b}_1^{'}, \mathbf{b}_2^{'}, \ldots, \mathbf{b}_L^{'}\}$ is the set of actual output vectors produced by the mapping network, then f minimizes the mean square error given by

$$E = \frac{1}{L} \sum_{l=1}^{L} \| \mathbf{b}_l - \mathbf{b}_l^{'} \|^2 . \tag{4.6}$$

An MLFFNN with at least two intermediate layers in addition to the input and output layers can perform a pattern mapping task [84, 87–91]. The additional layers are called the hidden layers. The neurons in these layers, called the hidden neurons, enable the network to learn complex tasks by extracting progressively more meaningful features from the input pattern vectors [86]. The input and output neurons for this task are linear units, while the hidden neurons are nonlinear units. The number of

units in the input and output layers is equal. Number of units in the hidden layers are more than the number of units in the input/output layers. For the mapping network, the mapping function f can be separated into f_1 and f_2 [92], so that

$$f(\cdot) \quad = \quad f_2(f_1(\cdot)), \tag{4.7}$$

where,

f_1 is the transformation in the mapping network from the input layer to the dimension expanding hidden layer, and

f_2 is the transformation from the dimension expanding hidden layer to the outer layer.

If the number of units in the input layer is n, and the number of units in the dimension expanding hidden layer is m (where $m > n$), f_1 transforms vectors in space R^n onto the space R^m. That is,

$$R^n \xrightarrow{f_1} R^m. \tag{4.8}$$

Similarly, f_2 transforms the higher dimensional space R^m back to the space R^n at the output.

$$R^m \xrightarrow{f_2} R^n. \tag{4.9}$$

Since $m > n$, f_1 is a dimension expansion process and f_2 is a dimension reduction process.

Dimension expansion is achieved by mapping the vectors in the input space onto a hypersurface captured by the set of weights in the network part of f_1. Dimension reduction is achieved by projecting the vectors in the hypersurface onto a subspace in the lower dimensional output space. The subspace is captured by the set of weights in the network part of f_2. The hypersurface and subspace are in general nonlinear, because of the nonlinear units in the hidden layers. Nonlinearity of the hidden neurons is necessary to provide generalization capability for the network. The number of hidden neurons must be large enough to form a complex decision region, but not so large enough that the weights cannot be reliably estimated from the available training data [93].

4.3.2.1 Structure of MLFFNN

The structure of the mapping network used in this work is *12L 24N 24N 12L*, where L refers to a linear unit and N to a nonlinear unit, the numbers represent the number of nodes in a layer (refer Fig. 4.3). The MLFFNN used for mapping consists of two hidden layers, because such a network can generate arbitrarily complex decision regions [84, 93]. When two hidden layers are used, the approximation improves compared to that when a single hidden layer is used [86]. The improvement in approximation is because the first hidden layer extracts the local features in the training data by partitioning the input space into regions. The second hidden layer then extracts the global features. A neuron in this layer combines the output of the neurons in the first hidden layer operating on a particular region of the input space, and thereby learns the global features for that region [86].

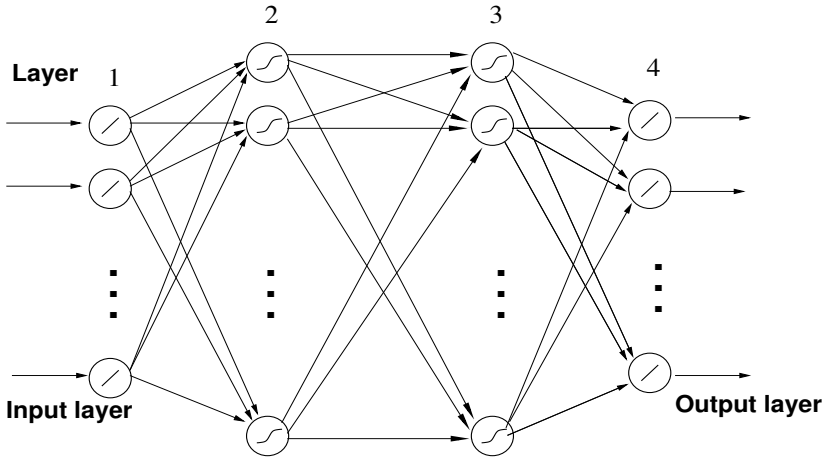

Figure 4.3: A 4 layer mapping neural network of size *12L 24N 24N 12L* where L refers to a linear unit and N to a nonlinear unit, the numbers represent the number of nodes in a layer.

4.3.2.2 Training the MLFFNN model

The mapping between the training pattern pairs involves iteratively determining the weights $\{w_{ij}\}$ of the network, such that $\mathbf{b}_l^{'}$ is equal (or nearly equal) to $\mathbf{b}_l$ for all the given L pattern pairs. The weights are determined by using the criterion that the total mean squared error between the desired output and the actual output is to be minimized.

To arrive at an optimum set of weights to capture the mapping implicit in the set of input-output pattern pairs, the conjugate gradient method [84] is preferred over the gradient descent method [84] in this work. This is because the conjugate gradient method converges much faster than the gradient descent method (refer Fig. 4.4).

In the conjugate gradient method, the increment in weight at the $(m+1)^{th}$ iteration is given by

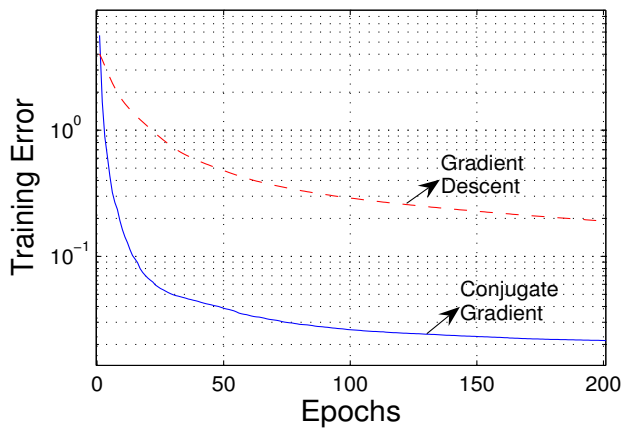

Figure 4.4: Training error curves for the gradient descent method and the conjugate descent method.

$$\Delta \mathbf{w} \quad = \quad \eta(m)\mathbf{d}(m), \tag{4.10}$$

where η is the learning rate parameter. The direction of increment $\mathbf{d}(m)$ in the weight is a linear combination of the current gradient vector and the previous direction of the increment in the weight [84]. That is,

$$\mathbf{d}(m) \quad = \quad -\mathbf{g}(m) + \alpha(m-1)\mathbf{d}(m-1), \tag{4.11}$$

where $\mathbf{g}(m)=\frac{\partial E}{\partial \mathbf{w}}$. The value of $\alpha(m)$ is obtained in terms of the gradient using the Fletcher-Reeves formula [84] given by

$$\alpha(m) \quad = \quad \frac{\mathbf{g}^T(m+1)\mathbf{g}(m+1)}{\mathbf{g}^T(m)\mathbf{g}(m)}. \tag{4.12}$$

The objective is to determine the value of η for which the error $E[\mathbf{w}(m) + \mathbf{d}(m)]$ is minimized for the given values of $\mathbf{w}(m)$ and $\mathbf{d}(m)$. Generally, the conjugate gradient

method converges much faster than the gradient descent method (refer Fig. 4.4), and hence is used in this work.

4.4 EXPERIMENTAL RESULTS

4.4.1 Training and testing the mapping network

The training and testing data are obtained from the same speaker because the mapping is speaker-dependent. The training data of each of the 10 speakers is of 5 minutes duration, and the testing utterances are of 20 sec duration. The simultaneously recorded speech signals from a throat microphone and a normal microphone are sampled at a rate of 8 kHz. LP analysis is performed on Hamming windowed speech frames, each of 20 msec duration. The overlap between adjacent frames is 5 msec. An LP order of $p = 8$ and the number of wLPCCs $q = 12$ are chosen empirically.

The wLPCCs of the TM speech and the NM speech form the input-output training pairs of the mapping network. Each training pattern is *preprocessed* so that its mean value, averaged over the entire training set, is close to zero. Each pattern (vector) is normalized so that the component values fall within the range $[-1, 1]$. This accelerates the training process of the network [86]. The training pattern pairs are presented to the network in the batch mode. The network is trained for 50 epochs. During testing, the wLPCCs of the TM speech are given as input to the mapping network. The network produces an output which are the estimated wLPCCs of the corresponding NM speech. The LP coefficients derived from these estimated wLPCCs (refer Section 4.3.1) are used to construct the all-pole synthesis filter.

4.4.2 Effect of mapping on various sound units

The effectiveness of the MLFFNN based mapping approach to map the TM spectra onto the NM spectra is analysed in this section. The LP spectra of the TM speech, the corresponding (desired) NM speech, and the estimated LP spectra for various sound units are shown in Fig. 4.5. It is clearly seen that the higher formants have a steep roll-off in the case of TM speech. In contrast, the spectral roll off in the enhanced speech is comparatively small. This shows that higher formants are emphasized in the enhanced speech. The LP spectra of the enhanced speech follows that of the NM speech for various sound units. For example, as seen in the figure, the LP spectra of the TM speech for the voiced stop consonants $/g/$ and $/d/$ resemble that of a vowel. This is due to the presence of formant-like structures during the closure phase. However, in the NM speech spectra, no such well-defined peaks are visible. In the enhanced speech too the LP spectra have no well-defined peaks. Further, the spectra of the enhanced speech and the NM speech appear similar. However, it is observed that the mapping is not learnt well in the case of fricatives. This is because of the random noise-like signal characteristic of fricatives. The LP spectra for a sequence of frames of the TM speech, the enhanced speech and the NM speech are shown in Fig. 4.6. It is seen that the spectra of the enhanced speech and the NM speech are similar. It is also seen that there are no spectral discontinuities in the spectra of the enhanced speech, indicating the effectiveness of the mapping approach.

4.4.3 Objective evaluation

The performance of this mapping technique is evaluated using the Itakura distance measure as the objective criterion. The Itakura distance is computed between two LP vectors. Since the LP vectors are related to the short-term spectra of the speech frames, this distance between the LP vectors indicates how similar the corresponding spectra are. The Itakura distance between two LP vectors, say $\mathbf{a}_k$ and $\mathbf{b}_k$, is given by [2]:

$$d_{ab}[\mathbf{a}_k, \mathbf{b}_k] = \frac{\mathbf{b}_k^T \tilde{\mathbf{R}}_{s_a} \mathbf{b}_k}{\mathbf{a}_k^T \tilde{\mathbf{R}}_{s_a} \mathbf{a}_k} \tag{4.13}$$

$$d_{ba}[\mathbf{a}_k, \mathbf{b}_k] = \frac{\mathbf{a}_k^T \tilde{\mathbf{R}}_{s_b} \mathbf{a}_k}{\mathbf{b}_k^T \tilde{\mathbf{R}}_{s_b} \mathbf{b}_k}, \tag{4.14}$$

where d_{ab} and d_{ba} are the asymmetric distances from $\mathbf{a}_k$ to $\mathbf{b}_k$, and vice versa, respectively. $\tilde{\mathbf{R}}_{s_a}$ and $\tilde{\mathbf{R}}_{s_b}$ are the autocorrelation functions of the speech frames corresponding to $\mathbf{a}_k$ and $\mathbf{b}_k$, respectively.

The Itakura measure is heavily influenced by spectral dissimilarity due to mismatch in formant locations, whereas errors in matching spectral valleys do not contribute heavily to the distance [2]. This is desirable, since the auditory system is more sensitive to the errors in formant locations and bandwidths than to the spectral valleys between peaks. Itakura distance is not a metric because it does not have the required property of symmetry. In order to achieve symmetry, d_{ab} and d_{ba} are combined as d=0.5($d_{ab}+d_{ba}$). This measure, in addition to symmetry, also has the property that if $\mathbf{a}_k$ and $\mathbf{b}_k$ are identical, the resulting distance is zero. Alternative spectral distortion metrics like the unweighted Euclidean distance are not appropriate here because the individual LP

coefficients in the vectors are highly correlated. The Euclidean distance is appropriate only when the representation of a vector is based upon an orthonormal basis set [2]. A weighted Euclidean distance, though appropriate, depends on finding a suitable decorrelating weighting matrix. Hence, the Itakura distance is used here.

The Itakura distance between the TM and the estimated LP coefficients, and the NM and estimated LP coefficients are computed for each frame. Fig. 4.7 shows the Itakura distance plots for two different utterances. It can be observed that the distance between the NM spectra and estimated spectra is very small when compared to the distance between the TM spectra and the estimated spectra. This shows that the estimated spectra are very close to the NM spectra. Thus the mapping network is able to capture the spectral correlation between the TM and the NM speech of a speaker.

4.5 SUMMARY

In this chapter, a method to capture the speaker-dependent functional relationship between the TM spectra and the NM spectra for achieving enhancement of the TM speech was proposed. The mapping of the spectra was modelled using a feedforward neural network. The stability of the all-pole synthesis filter was ensured by using the autocorrelation method to derive the LP coefficients from the estimated wLPCCs. The mapping performance was evaluated using the Itakura distance metric, as the Itakura measure is heavily influenced by spectral dissimilarity due to mismatch in formant locations. The advantage of this method was that, distortion due to spectral discontinuities between adjacent frames was not perceived in the enhanced speech, as discussed in the subjective studies in the next chapter (refer Section 5.7). This

chapter discussed the estimation of the spectral features of the NM speech, given the corresponding spectral features of the TM speech. The excitation source features are dealt with in the next chapter.

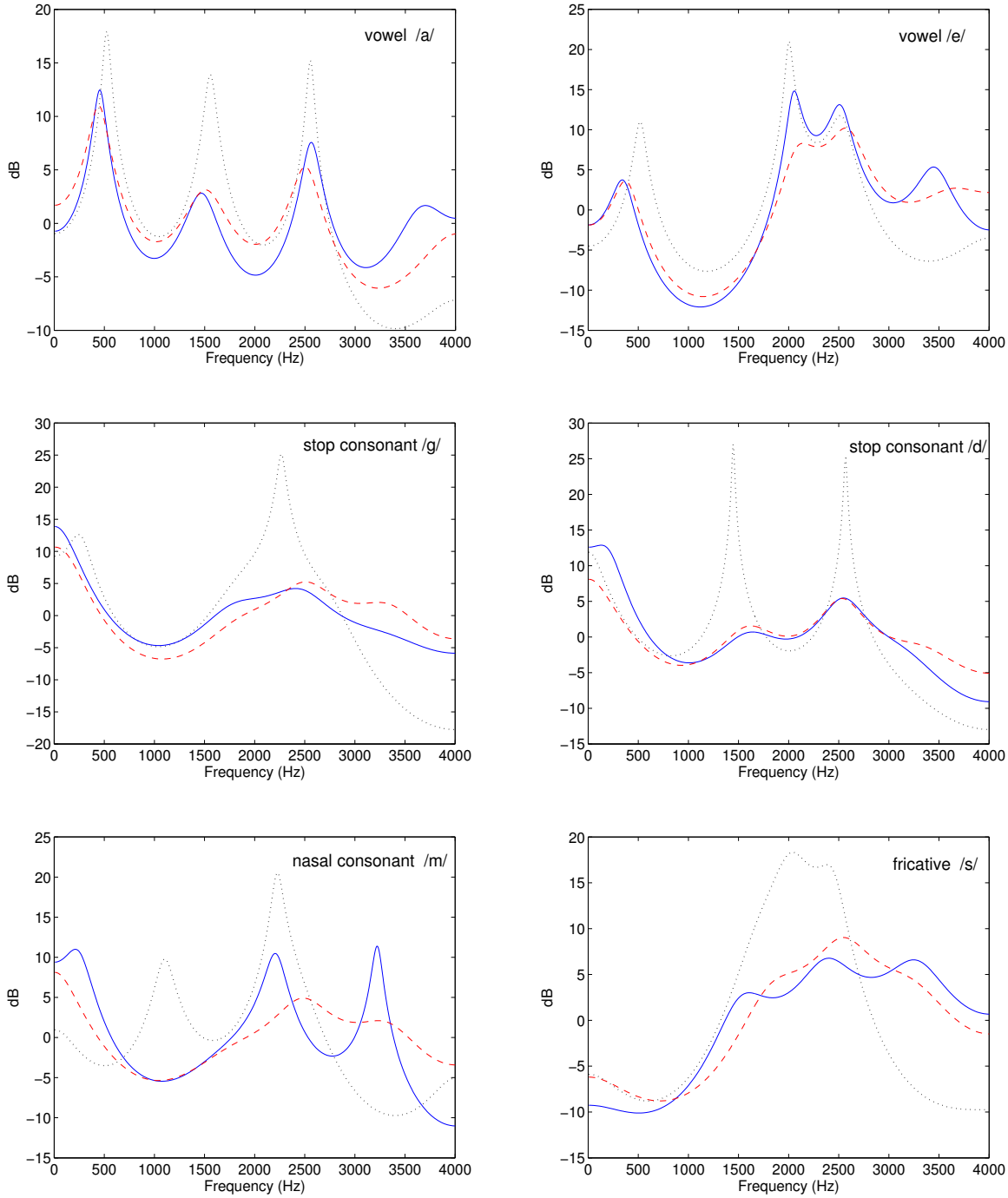

Figure 4.5: The LP spectra of the TM speech (shown as dotted line), the NM speech (shown as bold line), and the estimated LP spectra (shown as dashed line), for a frame of speech data for the sound units */a/*, */e/*, */g/*, */d/*, */m/*, and */s/*.

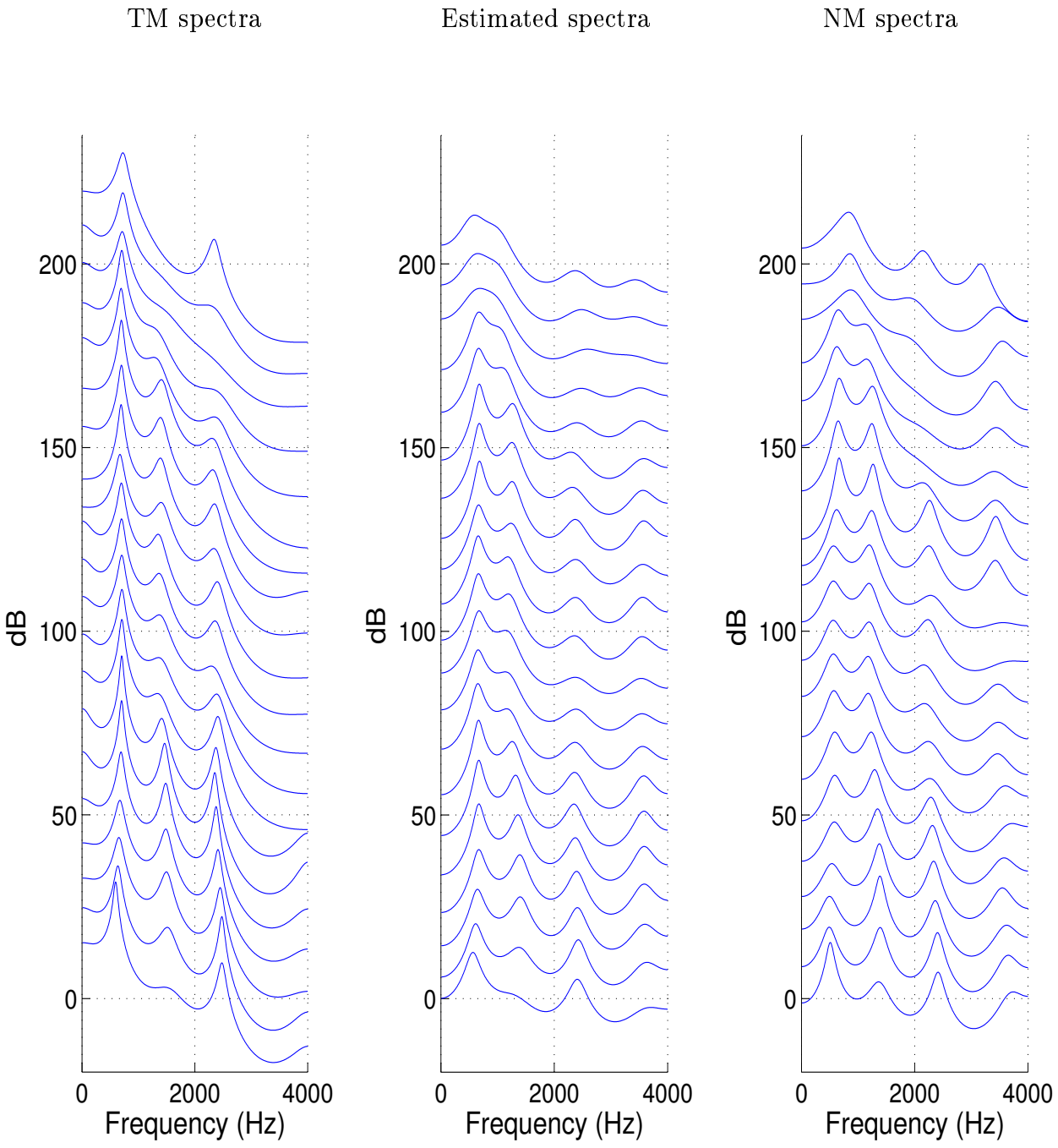

Figure 4.6: The LP spectra of the TM speech, the estimated LP spectra, and the LP spectra of the NM speech for a sequence of speech frames.

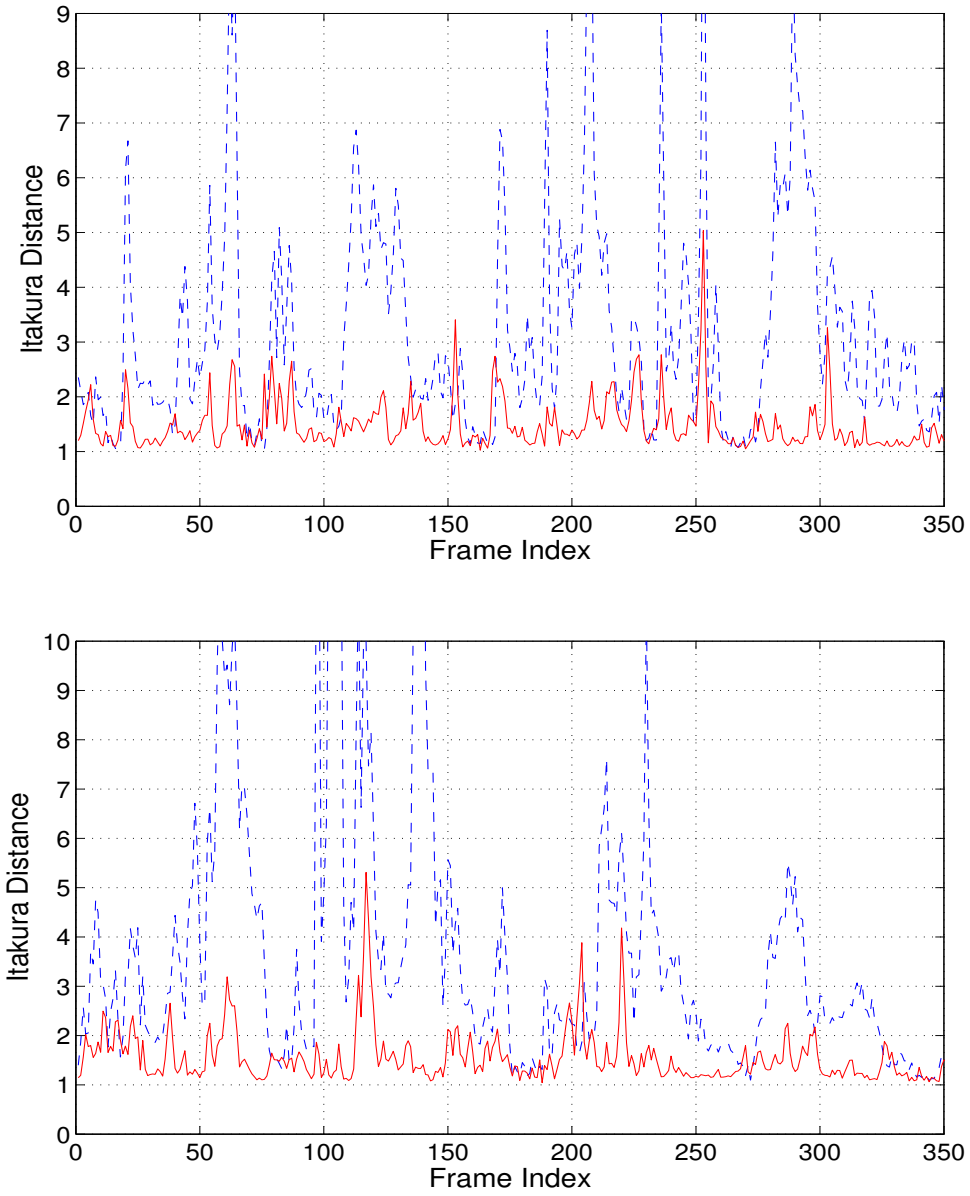

Figure 4.7: Itakura distance between the NM and TM spectra (dashed lines), and the NM and estimated spectra (solid lines) for two different speech utterances.

CHAPTER 5

ENHANCEMENT OF THROAT MICROPHONE SPEECH-RESIDUAL MODIFICATION

5.1 INTRODUCTION

The focus of the previous chapter has been on the first task in enhancing the TM speech, i.e., estimating the vocal tract characteristics of the NM speech. This involved mapping the spectral characteristics of the TM speech to the spectral characteristics of the NM speech. This chapter focuses on the second task in enhancing the TM speech, which involves estimating the excitation source characteristics of the NM speech. The LP residual of the TM speech varies from the LP residual of the NM speech for some of the voiced sounds like vowels, nasals and voiced stops (refer Section 3.7). To obtain an estimate of the NM residual signal, this variation needs to be compensated by suitably modifying the different voiced segments of the TM residual signal. The modified TM residual signal would be similar to the NM residual signal, which can then be used to excite the synthesis filter that is constructed using the estimated LP coefficients (refer Section 4.4) to obtain the enhanced speech.

This Chapter is organized as follows. The differences in the voiced segments of the LP residual of the TM and NM speech are discussed in Section 5.2. The features that help to distinguish between the voiced segments are discussed in Section 5.3. Mapping of these features of the TM and NM speech is discussed in Section 5.4. Section 5.5 explains the procedure to modify the TM residual to obtain an estimate of the NM

residual. The experimental results are discussed in Section 5.6, followed by a discussion on the subjective evaluation studies in Section 5.7. The work is summarized in Section 5.8.

5.2 TM RESIDUAL SIGNAL OF VOICED SEGMENTS

Differences exist in the LP residual signals of the voiced segments of the TM speech and the NM speech (refer Fig. 3.14). In the NM residual signal, the strength of the instants is comparatively high for the vowels and low for the voiced consonants. This relative emphasis of the vowel segments and deemphasis of the voiced consonant segments is referred to as *modulation* (of the strength of the instants) throughout this chapter. In the TM residual signal, such a modulation of the instants is not seen. The strength of the instants in the vowel segments is comparable to the strength of the instants in the voiced consonant segments.

The TM residual signal needs to be modified so as to achieve the modulation in the strength of the instants similar to that present in the NM residual signal. A simple scaling of the strength of the instants would not achieve the desired modulation because scaling would uniformly emphasize (or deemphasize) all the voiced segments. The modification should involve an emphasis of the strength of the instants in the vowel segments and a deemphasis of the instants in the voiced consonant segments. Such a modification would further improve the perceptual quality of the enhanced speech.

To modify the residual signal of the voiced segments of the TM speech, the broad phoneme category (e.g., vowels, nasals) of the speech segments needs to be identified.

A broad categorization is sufficient because the excitation source characteristics derived from the speech may be similar for the phonemes in each of the broad phoneme categories. For example, in the case of nasals, the effect of nasal coupling and loading of the vocal tract on the strength of the instants of excitation would be similar for different nasals. Once the broad phoneme categories are identified, the TM residual signal can then be suitably modified for each category of the voiced sounds. One approach to identify the broad categories of the voiced segments would involve building a speech recognizer (discussed in Section 6.3.2). This would require a large amount of speech data. The effectiveness of such an approach would depend on the performance of the recognizer.

An alternative approach to obtain an estimate of the NM residual signal would be to map the TM residual onto the NM residual. This approach will not yield the desired result because of the following reasons: The patterns used for mapping would be windowed (quasi-periodic) blocks of samples of the LP residual. It is known that the LP residual is sensitive to the position of the window. For example, if the windowing is done at regular intervals, then the variation in the positions of the instants among blocks would be large. This would result in a large variation in the training patterns. This variation in the positions of the instants and the variation in the dynamic range of the amplitude values of the samples of the LP residual will result in a poor mapping. Hence, in the mapped (estimated) NM residual signal, the periodicity of the pitch may not be maintained in adjacent frames of a voiced segment. This is because the positions of instants in a frame may not have co-located instants (otherwise present due to quasi-stationarity of pitch periodicity) in the succeeding frame belonging to

the same voiced sound. This would produce distortion in the synthesized speech. To reduce the variability in the training patterns, small blocks of the LP residual signal around the instants of voiced segments can be considered for mapping. However, this approach too will not produce the desired result since the open glottis regions are ignored in mapping. Hence, mapping the samples of the LP residual signal directly would not be effective for synthesis of speech.

As an alternative, the following method is used in this chapter. The aim is to modify the TM residual signal in order to approximate the envelope of the voiced segments of the NM residual signal. Hence, features that help to distinguish between the voiced segments in the speech signal are identified, as explained in the following section. These features derived from the TM signal are mapped onto the corresponding features derived from the NM signal using the mapping property of neural networks. The mapped features of the NM signal are used to suitably modify the TM residual signal of various voiced regions.

5.3 FEATURES FOR RESIDUAL MAPPING

An ideal feature for mapping would be the strength of the instants of the voiced segments. However, although the strength of the instants is visible in the time domain, deriving a robust parameter that represents the same is not easy. Hence, the following features are considered for mapping:

- Ratio of residual energy and signal energy
- Gross spectral features
- Log frame energy

These features derived from each analysis window of the speech signal are described in this section.

5.3.1 Ratio of residual energy and signal energy

The ratio of the residual energy and the signal energy, known as the normalized error (η), is related to the spectral dynamic range of the speech components. It decreases as the spectral dynamic range increases. Thus, in the NM speech, the normalized error, η_m, is small for vowels which have a large dynamic range, whereas it is large for the voiced consonants like nasals and voiced stops (refer Figs. 5.1 and 5.2, respectively) which have a comparatively lesser dynamic range. The normalized error is larger for unvoiced consonants like fricatives (than for voiced consonants), and silence regions.

In the TM speech, the normalized error, η_t, is generally (slightly) higher compared to η_m (refer the vowel segments in Figs. 5.1 and 5.2). This is because, the absence of some of the higher formants reduces the spectral dynamic range for vowels. For the voiced consonants, η_t is small, unlike η_m. This is because, for nasals the presence of oral resonances and the reduced effect of damping of the nasal cavity cause η_t to be small. For the voiced stops too, η_t is small, as the dynamic range is large due to the well-defined formant-like structures associated with the closure phase (refer Fig. 3.4). Thus, differences exist in η_t and η_m for the vowel segments as well as the voiced consonant segments. A parameter that can represent this difference is the ratio of η_t and η_m, given by

$$R_\eta = \frac{\eta_t}{\eta_m}. \tag{5.1}$$

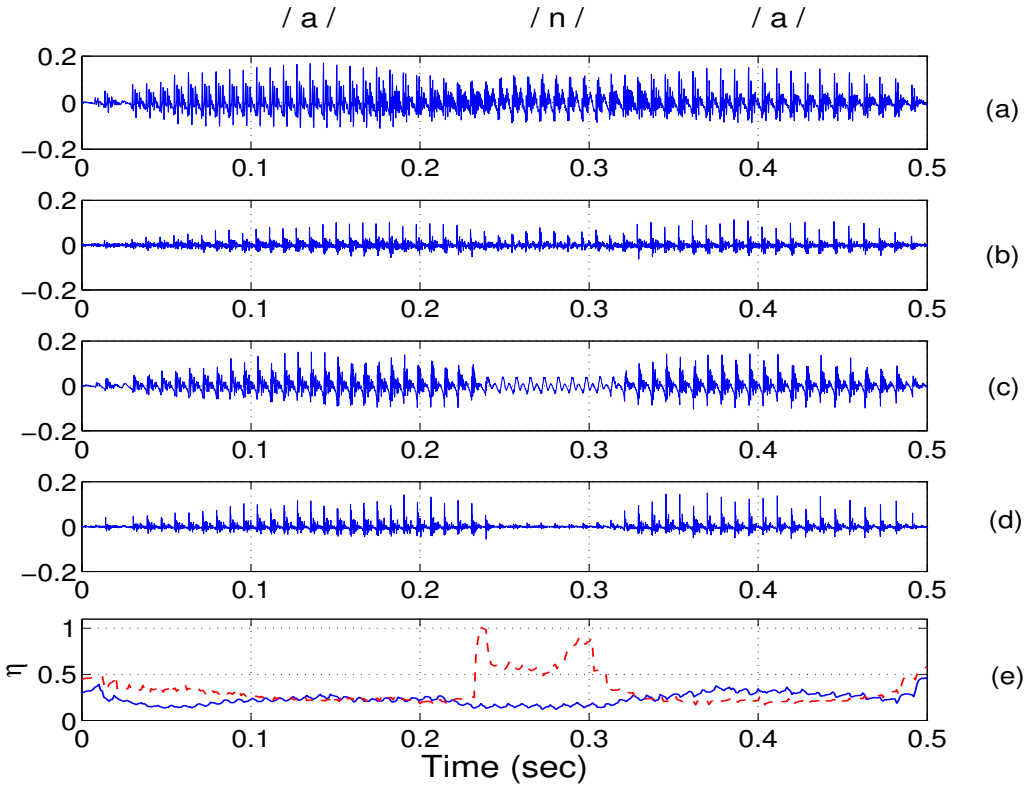

Figure 5.1: (a) TM speech and (b) its LP residual, (c) NM speech and (d) its LP residual, and (e) η_t (solid line) and η_m (dashed line), for the syllable */ana/*

The difference between the values of R_η for the vowels and voiced consonants is illustrated in Fig. 5.3. It is seen from Fig. 5.3 (f) that, mostly, $R_\eta \leq 1$ for the voiced consonants (/m/ and /b/), and $R_\eta \geq 1$ for vowels. This shows that R_η could be used as a parameter to distinguish different types of voiced segments in the TM speech. Since only the TM speech and hence η_t is available, η_m can be estimated from the mapping, in order to compute R_η.

5.3.2 Gross spectra of voiced sound categories

Spectral features vary for different voiced sound units. The spectral variation among the sound units within each broad phoneme category is high, when a higher order LP

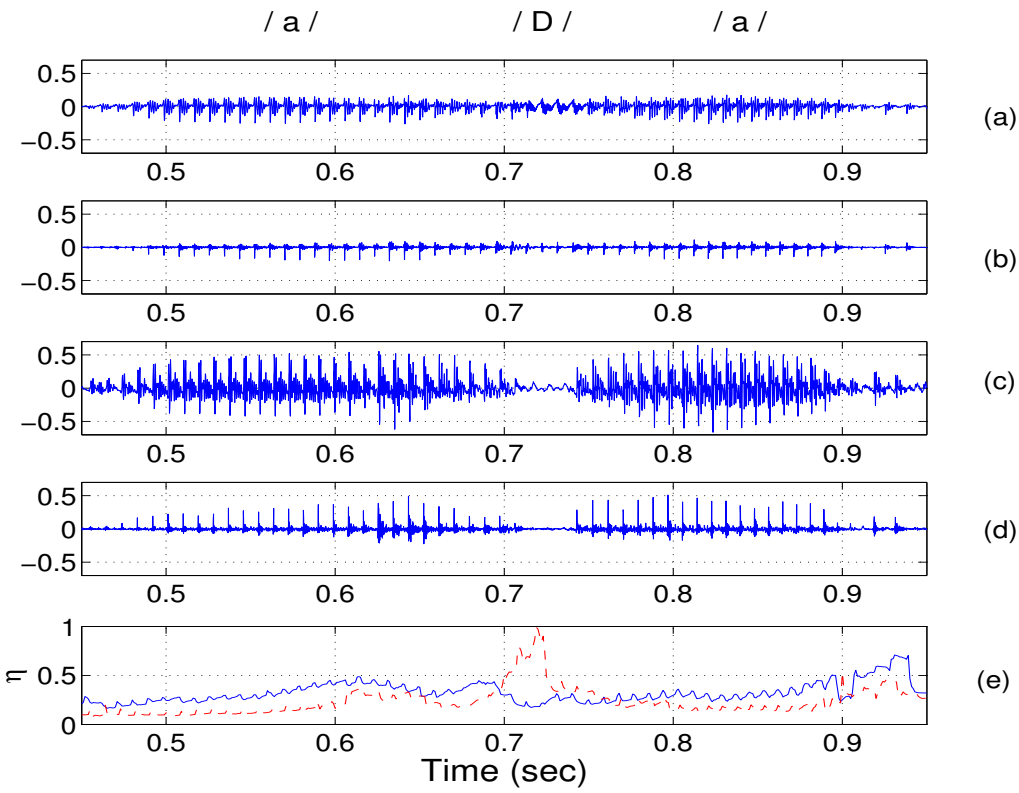

Figure 5.2: (a) TM speech and (b) its LP residual, (c) NM speech and (d) its LP residual, and (e) η_t (solid line) and η_m (dashed line), for the syllable $/aDa/$

analysis is used. As the number of formants is determined by the order of the LP analysis (one complex pole pair accounts for one formant), a higher order LP analysis results in more number of formants. The locations of the formants vary among the sound units within each broad phoneme category depending on the place of articulation of that sound. When a lower order LP analysis is used, there is a smoothing of the spectrum as the resonant bandwidths are broadened and closely spaced resonances are smeared. As the finer spectral differences are lost, the lower order spectral envelopes tend to represent the broad category of a sound unit. If a higher order LP spectrum is used for mapping, then the spectrum may dominate the mapping, which is undesirable.

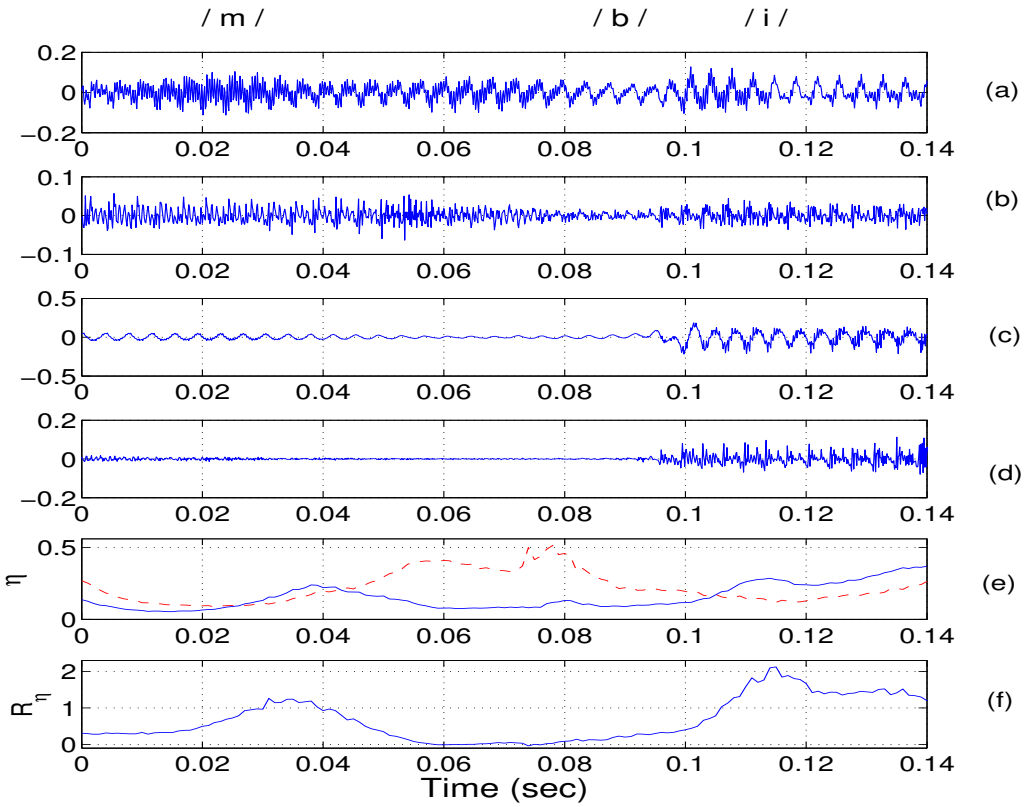

Figure 5.3: (a) TM speech and (b) its LP residual, (c) NM speech and (d) its LP residual, (e) η_t (solid line) and η_m (dashed line), and (f) R_η for the syllable */mbi/*.

In contrast, if a lower order LP spectrum is used, the mapping of the broad phoneme categories is expected to be effective.

In Figs. 5.4 and 5.5, the spectra of various voiced sounds of the TM speech obtained using a lower order LP analysis ($p = 3$) are compared with the corresponding spectra obtained using a higher order LP analysis ($p = 8$). The higher order spectral envelopes show the finer details in the resonant structure of the vocal tract system that distinguish each of the sounds within a category. In contrast, in the lower order spectral envelopes the finer details in the resonant structures are lost. Hence, a lower

order LP spectrum is sufficient to represent the broad category of the voiced segment.

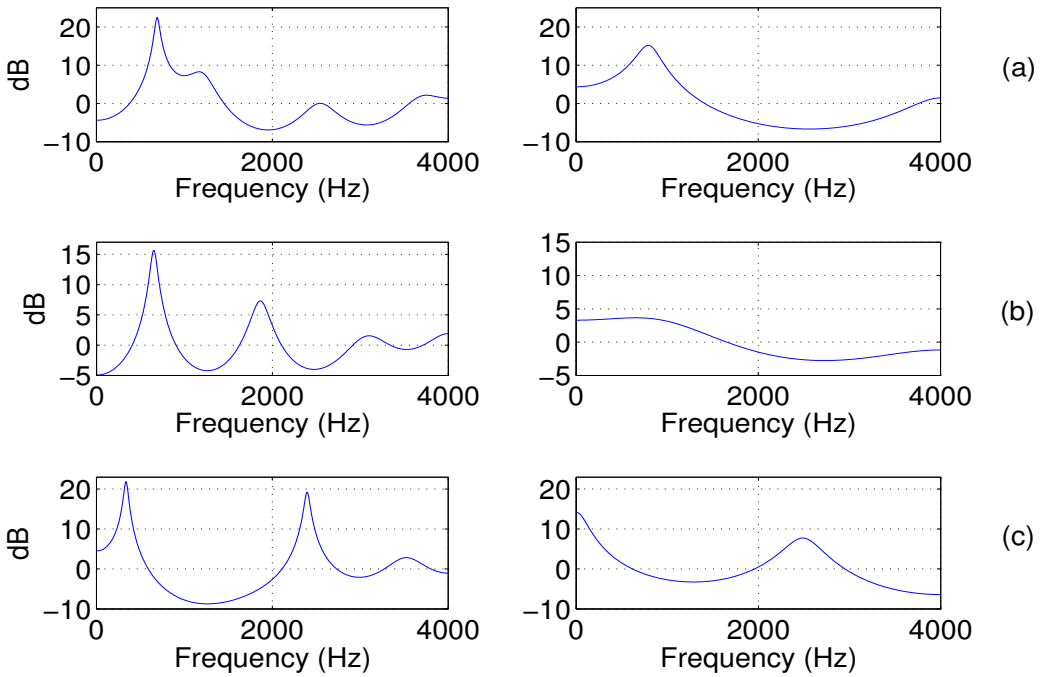

Figure 5.4: A comparison of the LP spectra derived from LP analysis with orders 8 and 3 for the vowels (a) /*a*/, (b) /*e*/, and (c) /*i*/ in the TM speech.

5.3.3 Log frame energy

The log frame energy (G) of the speech signal s, which is a measure of the intensity of the signal, differs in the TM and NM speech for various sound units. The log frame energy is computed for each window of N samples as $G = \log\frac{1}{N}\sum_{i=1}^{N} s_i^2$, where s_i is the ith sample in the window. Figs. 5.6 and 5.7 show the TM and NM speech signal waveform and log frame energy contours for two segments of speech. The log frame energy, G_m, of the NM speech signal is comparatively high for the vowel segments, and low for voiced consonants. However, in the TM speech, the log frame energy, G_t, is comparable in the vowel and voiced consonant segments. The log frame energy

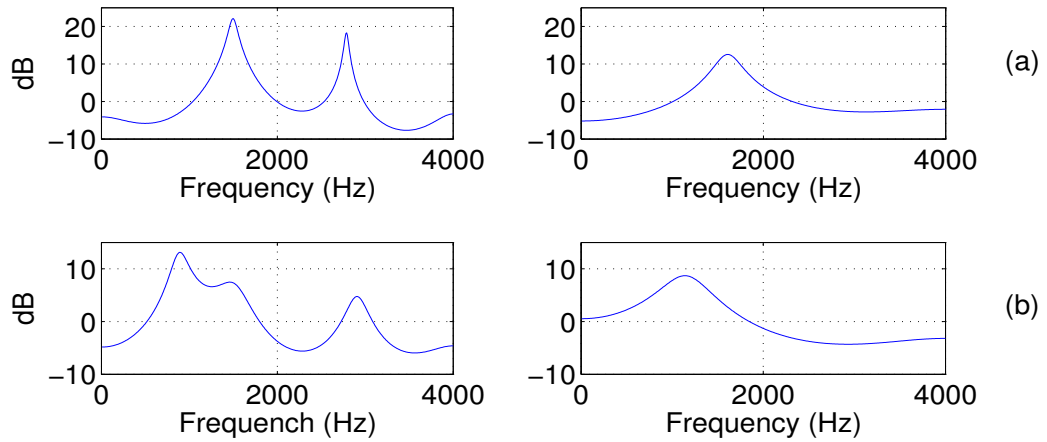

Figure 5.5: A comparison of the LP spectra derived from LP analysis with orders 8 and 3 for the nasals (a) /n/, and (b) /m/ in the TM speech.

of the NM speech that is estimated from the mapping could be used to control the gain in the speech that is synthesized using the modified residual. Prior to mapping, G is smoothed to reduce fluctuations. The logarithm of the frame energy is used as it compresses the dynamic range of the energy values, and makes the mapping less sensitive to the small variations in the input.

5.4 RESIDUAL MAPPING USING MLFFNN

The features for mapping are extracted from analysis windows (or frames) of 20 ms each, with an overlap of 1 ms between adjacent frames. The feature vector of the ith frame is 11-dimensional, comprising of (a) 5-dimensional wLPCCs $\{\mathbf{c_i}\}$ (derived using an LP analysis of order 3), (b) the normalized error of that frame (η_i) and the adjacent frames (η_{i-1} and η_{i+1}), and (c) the log frame energies of that frame (G_i) and the adjacent frames (G_{i-1} and G_{i+1}). The normalized errors and the log frame energies of the adjacent frames are used along with that of the current frame so as to capture

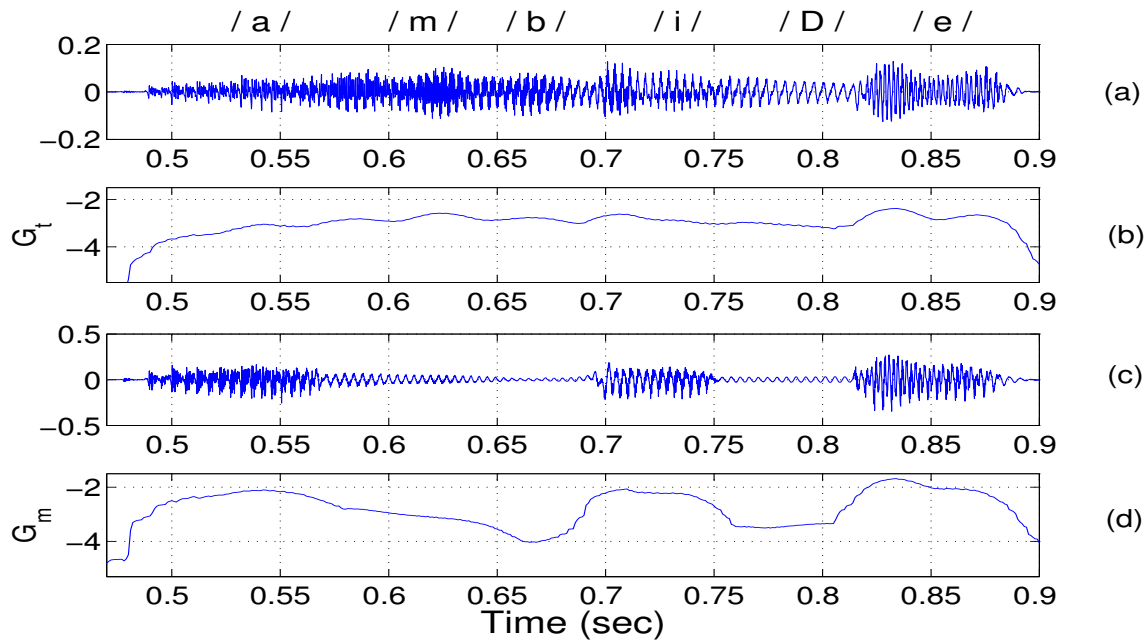

Figure 5.6: The TM and NM speech ((a) and (c), respectively), and their corresponding log frame energy contours ((b) and (d), respectively) for the syllable *ambide*, which is part of the word *ambidextrous*.

the constraint in the temporal sequence in mapping.

The mapping neural network consists of two hidden layers, apart from the input and output layers. The size of the network is $11L - 22N - 22N - 11L$. The network is trained using the features derived from 5 minutes of speech data recorded simultaneously using the throat microphone and the normal microphone. The features from the TM data ($\{\mathbf{c_{t_i}}\}, \{\eta_{t_{i-1}}, \eta_{t_i}, \eta_{t_{i+1}}\}, \{G_{t_{i-1}}, G_{t_i}, G_{t_{i+1}}\}$) form the input, and the corresponding features from the NM data ($\{\mathbf{c_{m_i}}\}, \{\eta_{m_{i-1}}, \eta_{m_i}, \eta_{m_{i+1}}\}, \{G_{m_{i-1}}, G_{m_i}, G_{m_{i+1}}\}$) form the target output for training. The speaker-specific mapping uses the conjugate gradient method for learning (as explained in Section 4.3.2.2). Speaker-specific mapping is preferred to speaker-independent mapping so as to overcome the differences that may exist in the values of R_η and G among speakers. The differences may result

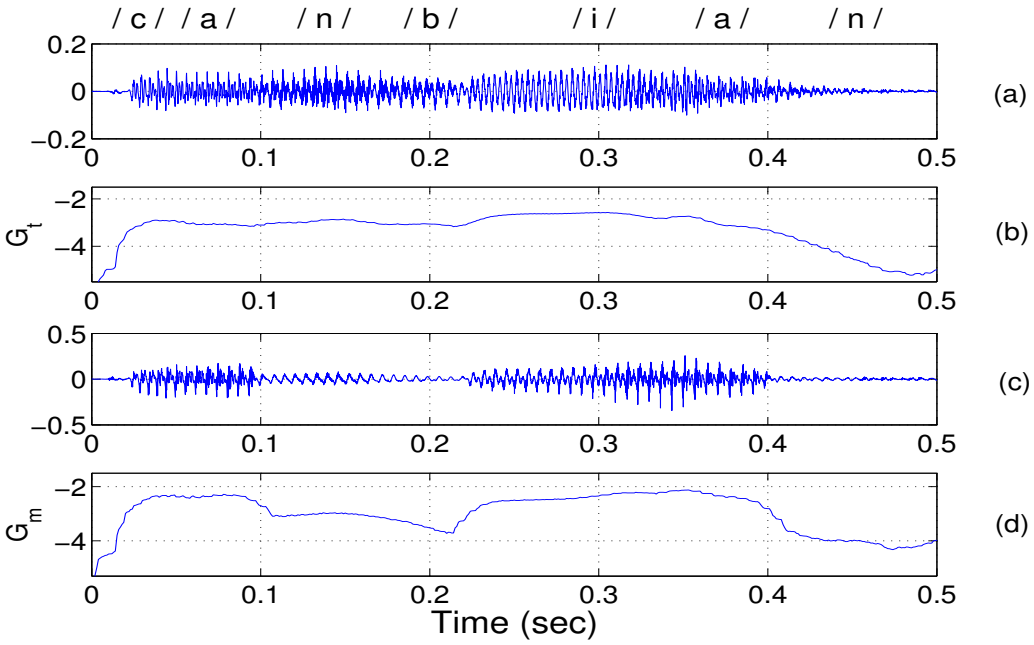

Figure 5.7: The TM and NM speech ((a) and (c), respectively), and their corresponding log frame energy contours ((b) and (d), respectively) for the utterance *can be an*.

in undue emphasis or deemphasis of the speech segments in the TM residual signal, causing perceivable distortion in the synthesized speech. The estimate of the normalized error, $\hat{\eta}_m$ and log frame energy, $\hat{G}_m$ of the NM speech are used to modify the TM residual signal, as explained in the next section.

5.5 MODIFICATION OF TM RESIDUAL SIGNAL USING MAPPED FEATURES

In order to modify the strengths of the instants of the voiced segments of the TM residual signal, the instants of glottal closure have to be identified. The instants of glottal closure are not identified directly from the LP residual, though the LP residual contains information pertaining to the excitation. This is because of the following reasons [94]: The LP analysis of the speech signal assumes an all-pole model. The all-

pole model implicitly assumes a minimum-phase characteristic for the speech signal. If this assumption is invalid, then the phase response of the vocal tract system is not compensated exactly by the inverse filter. The effect of the uncompensated phase on the LP residual is not known. Also, the inverse filter does not compensate for the zeros which may be introduced due to the finite duration of the glottal pulse or nasal coupling. These factors cause multiple peaks of either polarity to occur around the instants of glottal closure in the LP residual. This makes it difficult to estimate the instants from the LP residual.

Instants of glottal closure could be better estimated if impulse-like signals could be obtained at the instants of significant excitation. The Hilbert envelope of the LP residual is a close approximation to obtain such impulse-like signals around the instants [94, 95]. The Hilbert envelope of the LP residual signal is a positive function, giving the envelope of the signal [96]. The Hilbert envelope $h(n)$ of the LP residual $r(n)$ is given by [95, 96]

$$h(n) = \sqrt{r^2(n) + {r_h}^2(n)}, \tag{5.2}$$

where $r_h(n)$ is the Hilbert transform of $r(n)$, and is given by

$$r_h(n) = IDFT[R_h(\omega)], \tag{5.3}$$

where

$$R_h(\omega) = \begin{cases} -jR(\omega), & 0 \leq w < \pi \\ jR(\omega), & -\pi < w < 0, \end{cases} \tag{5.4}$$

where IDFT is the inverse discrete Fourier transform, and $R(\omega)$ is the discrete Fourier transform of $r(n)$. The peaks in the Hilbert envelope indicate the locations of the instants (refer Fig. 5.8). Using the Hilbert envelope resolves the ambiguities in identifying the instants directly from the residual signal. Hence the Hilbert envelope of the TM residual signal is used for residual modification as explained below.

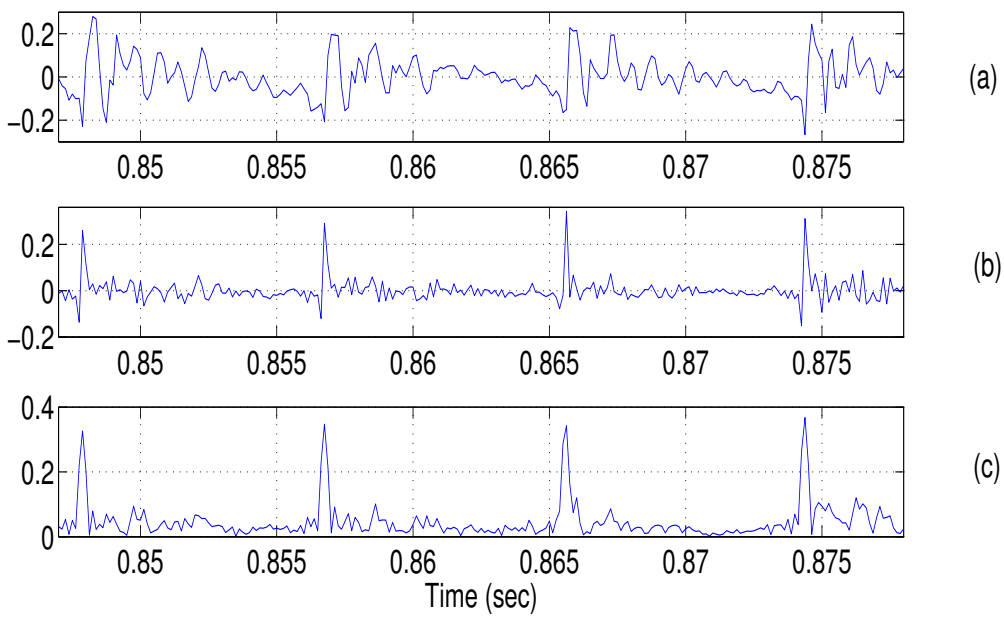

Figure 5.8: (a) A speech signal, (b) its LP residual, and (c) Hilbert envelope of the LP residual for a segment of vowel /a/. The quasi-periodic peaks indicate the instants of excitation.

The Hilbert envelope, $h_t(n)$ of the LP residual of the TM speech ($s_t(n)$) to be enhanced is scaled using the running mean to normalize the signal, and is given by

$$h_T(n) = \frac{h_t^2(n)}{\frac{1}{N}\sum_{i=n-\frac{N}{2}}^{n+\frac{N}{2}} h_t(i)}, \tag{5.5}$$

where N is the length of a small window (typically 5 ms) around $h_t(n)$. The strength of the instants of $h_T(n)$ is modulated by scaling $h_T(n)$ with the estimated parameter $\hat{R}_\eta(n)$.

$$\hat{h}_m(n) = h_T(n)^{\frac{1}{\hat{R}_\eta(n)}}, \tag{5.6}$$

where

$$\hat{R}_\eta = \frac{\eta_t}{\hat{\eta}_m}. \tag{5.7}$$

The modulated Hilbert envelope $\hat{h}_m(n)$ is an estimate of the corresponding Hilbert envelope ($h_m(n)$) of the NM residual signal. The modified residual signal, $\hat{r}_m(n)$, is obtained using $\hat{h}_m(n)$ and the cosine of the phase of the analytic signal corresponding to the TM residual signal as follows.

The analytic signal $r_a(n)$ corresponding to the LP residual signal $r(n)$ is given by

$$r_a(n) = r(n) + jr_h(n). \tag{5.8}$$

The magnitude of the analytic signal, $|r_a(n)|$, is the Hilbert envelope of $r(n)$. The cosine of the phase of the analytic signal is given by

$$\cos(\theta(n)) = \frac{\mathbf{Re}(r_a(n))}{\mid r_a(n) \mid} = \frac{r(n)}{h(n)}. \tag{5.9}$$

The estimated LP residual of the NM speech, $\hat{r}_m(n)$, is given by

$$\hat{r}_m(n) \quad = \quad \hat{h}_m(n)\cos(\theta_t(n)). \tag{5.10}$$

The estimated LP residual, $\hat{r}_m(n)$, is used to excite the all-pole synthesis filter, $H(z) = \frac{1}{1-\sum_{i=1}^{p} \hat{a}_i z^{-1}}$, to obtain the enhanced speech signal $s(n)$. Here $\{\hat{a}_0, \hat{a}_1, \cdots, \hat{a}_p\}$ are the LP coefficients derived from the estimated wLPCCs (refer Section 4.3.1). The gain in the speech signal $s(n)$ is modified using the estimated log frame energy, $\hat{G}_m(n)$, to obtain the enhanced speech as

$$\hat{s}_m(n) = s(n)e^{\hat{G}_m(n)}. \tag{5.11}$$

5.6 EXPERIMENTAL RESULTS

Fig. 5.9 shows the modification sequence of the Hilbert envelope ($h_t(n)$) of the TM residual signal, for the utterance *ambidex* (in the word *ambidextrous*). The strength of $h_t(n)$ (refer Fig. 5.9 (a)) is comparable in the vowel (/a/) and voiced consonant (/m/ and /b/) segments. The removal of the trend in $h_t(n)$ results in an increase in the strength of the signal ($h_T(n)$) in most of the voiced (vowel and voiced consonants) segments, as seen in Fig 5.9 (b). The signal $h_T(n)$ is modified using the parameter $\hat{R}_\eta(n)$, which is shown in Fig. 5.9 (c). It is seen that the estimate $\hat{R}_\eta(n)$ is a close approximate of $R_\eta(n)$ for most of the sound units. The modified Hilbert envelope, $\hat{h}_m(n)$ is shown in Fig. 5.9 (d). It is seen that the vowel (/a/ and /e/) regions are emphasized relative to the voiced consonant (/m/, /b/ and /D/) regions. Also, the

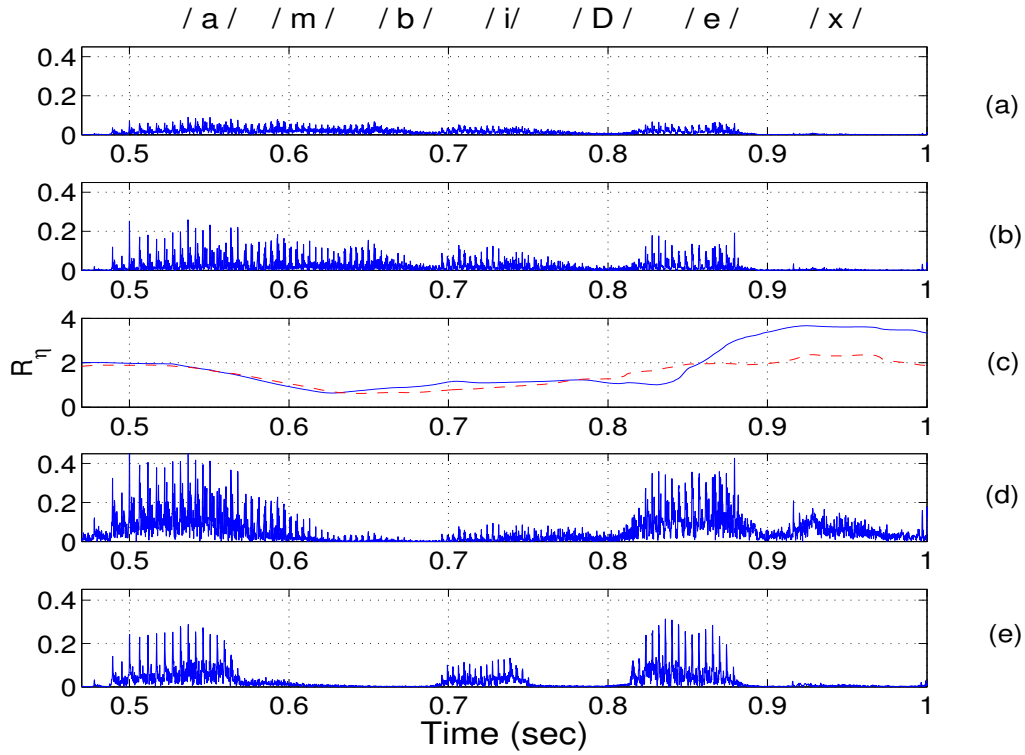

Figure 5.9: The mapping-based modification of the Hilbert envelope of the TM residual signal. (a) $h_t(n)$, (b) $h_T(n)$, (c) $\hat{R}_\eta(n)$ (dashed line) and $R_\eta(n)$ (solid line), (d) $\hat{h}_m(n)$, and (e) h_m for the speech segment */ambidex/* in the word *ambidextrous*.

modulation of the strength of the instants in $\hat{h}_m(n)$ is comparable with that of $h_m(n)$, shown in Fig. 5.9 (e).

The estimated LP residual signal, $\hat{r}_m(n)$, is shown in Fig. 5.10 (b). The LP residual signals of the TM and NM speech are shown in Figs. 5.10 (a) and 5.10 (c), respectively. It is seen that the strength of the TM residual signal ($r_t(n)$) is similar for most of the voiced segments. In contrast, the strength of $\hat{r}_m(n)$ is relatively large in the vowel segments, compared to the voiced consonant segments. This relative difference in the strength between the vowel and voiced consonant segments is similar to the difference in the NM residual signal ($r_m(n)$).

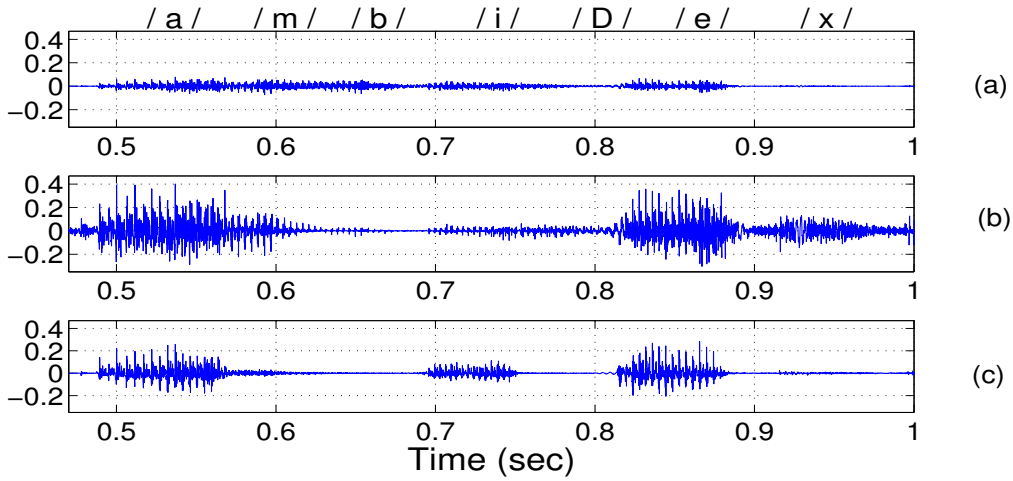

Figure 5.10: LP residual signal (a) derived from TM speech, $r_t(n)$, (b) estimated, $\hat{r}_m(n)$, and (c) derived from NM speech, $r_m(n)$ for the speech segment */ambidex/* in the word *ambidextrous*.

The speech signal, $s(n)$, synthesized using $\hat{r}_m(n)$ is shown in Fig. 5.11 (b). The estimated gain parameter $\hat{G}_m(n)$ is shown in Fig. 5.11 (c). The gain in $s(n)$ is modified using $\hat{G}_m(n)$ to obtain $\hat{s}_m(n)$, as shown in Fig. 5.11 (d). The overall envelope of the enhanced speech signal $\hat{s}_m(n)$ is similar to the signal envelope of $s_m(n)$ (NM speech shown in Fig. 5.11 (e)).

The gain modification using $\hat{G}_m(n)$ is seen to compensate for the large emphasis/deemphasis caused by $\hat{R}_\eta(n)$, as illustrated in Figs. 5.12 through 5.14. Fig. 5.12 shows the modification sequence of $h_t(n)$ for the utterance *can be an*. The modified Hilbert envelope signal, $\hat{h}_m(n)$, in Fig. 5.12 (d) shows a large deemphasis of the voiced segment */bi/*, and a large emphasis of the silence region (following the last phoneme (*/n/*) of the utterance). The undue emphasis of the silence region is because $R_\eta(n)$ is more appropriate for the voiced segments than for the unvoiced segments. This large deemphasis/emphasis is also observed in $\hat{r}_m(n)$ and $s(n)$ (refer Figs. 5.13 and 5.14,

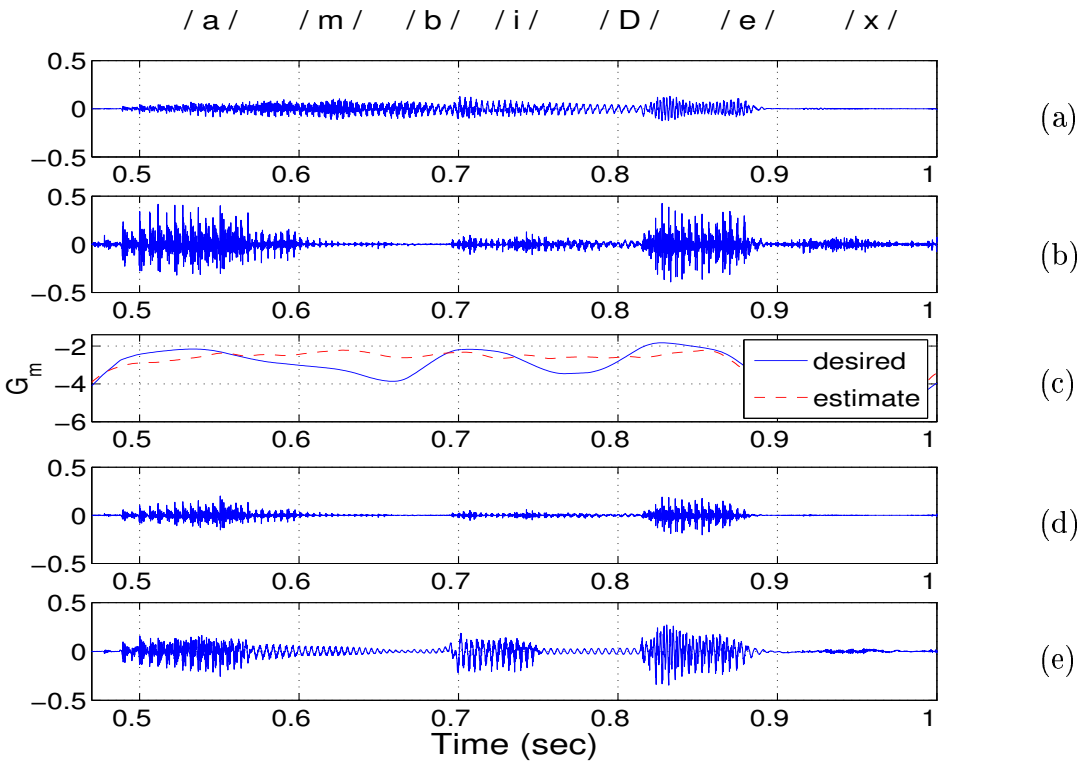

Figure 5.11: (a) TM speech, $s_t(n)$, (b) speech $s(n)$ synthesized using $\hat{r}_m(n)$, (c) log frame energy, $\hat{G}_m(n)$ (dashed line) and G_m (solid line), (d) gain-modified speech, $\hat{s}_m(n)$, and (e) NM speech, $s_m(n)$ for the speech segment */ambidex/* in the word *ambidextrous*.

respectively).

When the gain in $s(n)$ is modified using $\hat{G}_m(n)$ (shown in Fig. 5.14 (c)), the resulting enhanced speech, $\hat{s}_m(n)$ (refer Fig. 5.14 (d)), shows an increase in the amplitude of the signal in the segment */bi/*. Similarly, in the silence segment the gain modification suppresses the emphasis caused by $\hat{R}_\eta(n)$. This is because $\hat{G}_m(n)$ is low for the unvoiced/silence regions.

In general, the parameter $\hat{R}_\eta(n)$ emphasizes the instants in the vowel regions, and deemphasizes the instants in the voiced consonant regions. The gain parameter $\hat{G}_m(n)$

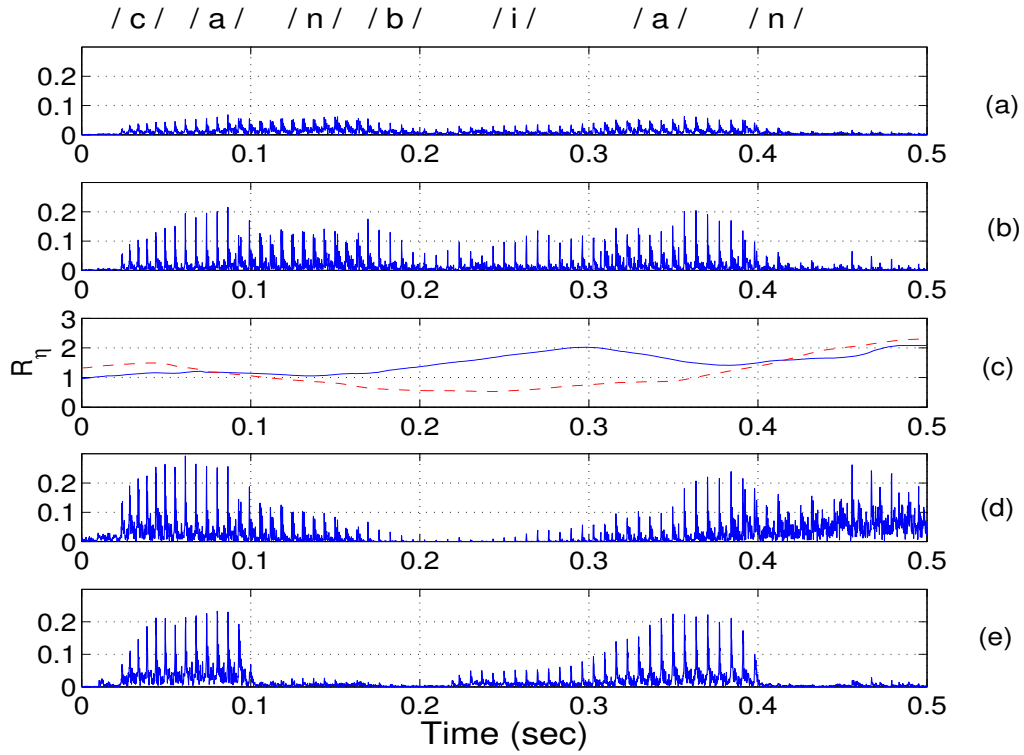

Figure 5.12: The mapping-based modification of the Hilbert envelope of the TM residual signal. (a) $h_t(n)$, (b) $h_T(n)$, (c) $\hat{R}_\eta(n)$ (dashed line) and $R_\eta(n)$ (solid line), (d) $\hat{h}_m(n)$, and (e) h_m for the speech segment *can be an*.

compensates for the large emphasis/deemphasis caused by $\hat{R}_\eta(n)$, both in the voiced as well as unvoiced/silence regions. This modification of the TM residual results in obtaining a close approximate of the envelope of the NM residual signal. However, the results show that a further deemphasis of the nasal segments is desired in order to achieve a better estimate of the envelope of the NM residual signal.

5.7 SUBJECTIVE EVALUATION

In order to evaluate the subjective quality of the enhanced signal, the Comparison Mean Opinion Score (CMOS) test is executed [4]. The CMOS subjective test is chosen

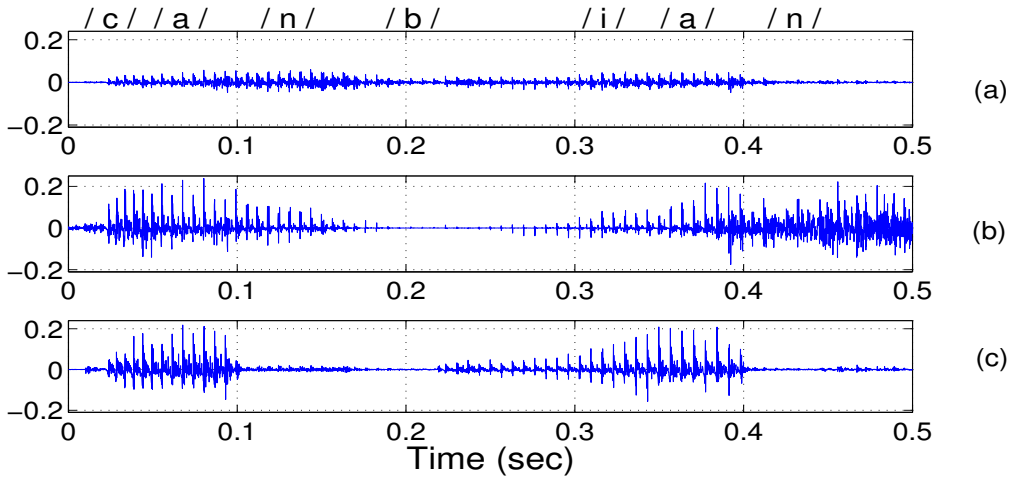

Figure 5.13: LP residual signal (a) derived from TM speech, $r_t(n)$, (b) estimated, $\hat{r}_m(n)$, and (c) derived from NM speech, $r_m(n)$ for the speech segment *can be an*.

since it is of interest to compare the perceptual quality of the enhanced speech with that of the TM speech to assess the improvement in naturalness provided by the enhancement technique.

The listeners for this test consist of 20 graduate students who volunteered for the task. Listeners are presented with a pair of speech samples on each trial. The pair consists of an unprocessed speech sample and the corresponding processed sample. The order of the samples presented for each trial is chosen at random. Listeners judge the quality of the second sample relative to the first using a 7-point scale given in Table 5.1. Prior to the evaluation, listeners are familiarized with the TM and enhanced speech signals different from those used for evaluation. The CMOS is computed for the (unprocessed, processed) order of presentation of speech.

The results of the evaluation are shown in Fig. 5.15 in the form of histograms. A comparison of the TM and NM speech shows a clear preference for the NM speech

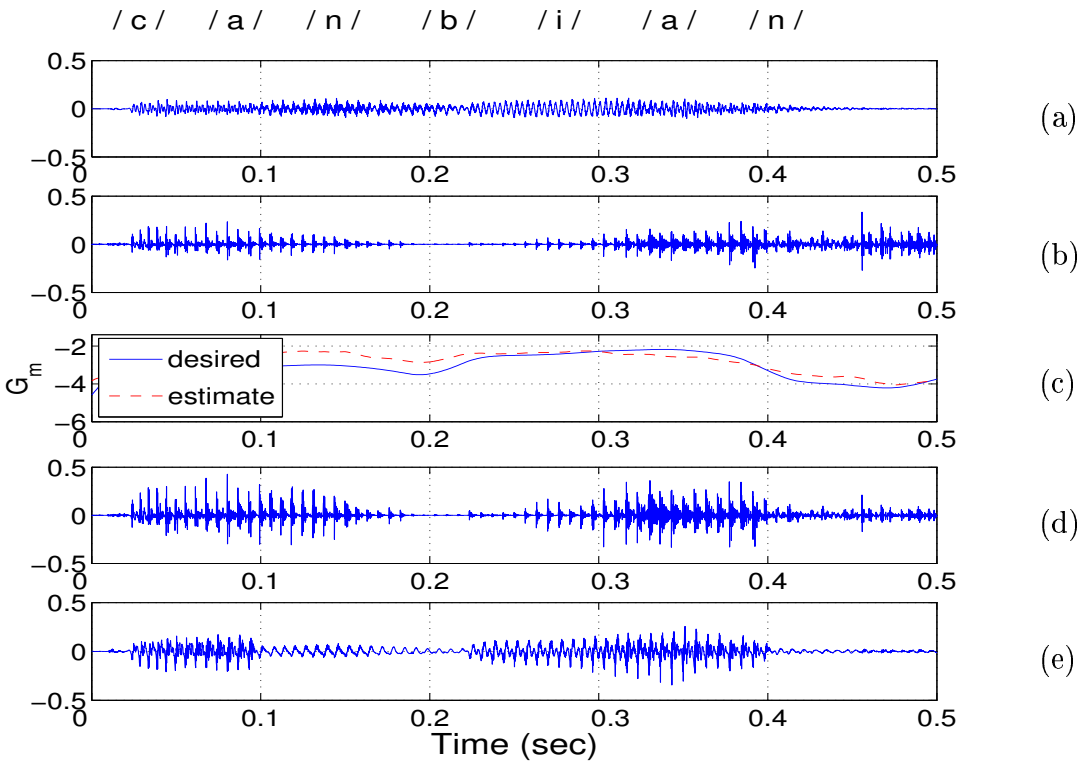

Figure 5.14: (a) TM speech, $s_t(n)$, (b) speech $s(n)$ synthesized using $\hat{r}_m(n)$, (c) log frame energy, $\hat{G}_m(n)$ (dashed line) and G_m (solid line), (d) gain-modified speech, $\hat{s}_m(n)$, and (e) NM speech, $s_m(n)$ for the speech segment *can be an*.

($CMOS = 2$). To evaluate the progress in the enhancement technique, the TM speech ($s_t(n)$) is compared with the speech ($s_l(n)$) synthesized with the estimated LP coefficients and the unmodified TM residual. The partially enhanced speech, $s_l(n)$, has a higher rating compared to the TM speech ($CMOS = 1.14$, between sightly better and better than the TM speech). The effect of modifying the TM residual signal is assessed by comparing $s_l(n)$ with the speech ($\hat{s}_m(n)$) synthesized using the estimated LP coefficients and the modified TM residual. The $\hat{s}_m(n)$ has only a marginally higher rating compared to $s_l(n)$ ($CMOS = 0.32$). A few listeners have rated $\hat{s}_m(n)$ as 'slightly

Table 5.1: A 7-point rating used in the Comparison Mean Opinion Score (CMOS) test to judge the quality of the second speech sample relative to that of the first speech sample.

Rating	Speech Quality
3	Much Better
2	Better
1	Slightly Better
0	About the Same
-1	Slightly Worse
-2	Worse
-3	Much Worse

degraded' compared to $s_l(n)$. This is because the modification of the TM residual causes undue deemphasis of a few voiced segments. A comparison of $\hat{s}_m(n)$ with the NM speech ($s_m(n)$) shows the preference for the NM speech ($CMOS = -1.1$). This indicates that the existing technique needs to be explored further to achieve better results. The speech signals are available for listening at the site `http://speech.cs.iitm.ernet.in/Main/result/MappingThroatToNormalSpeech/`.

5.8 SUMMARY

In this chapter, the residual of the voiced segments of the TM speech was modified to approximate the envelope of the residual of the voiced segments of the NM speech. The ratio of the normalized error of the TM speech and the estimated normalized error of

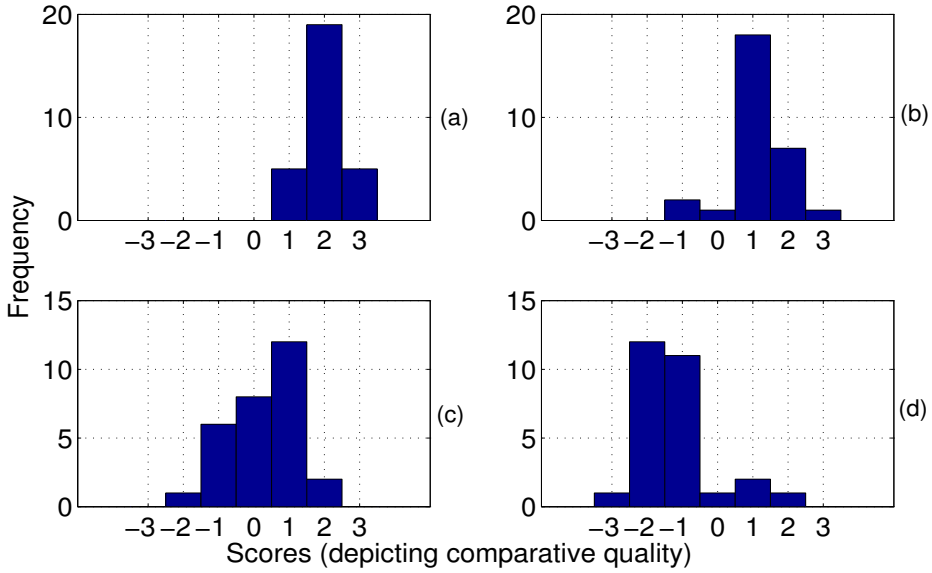

Figure 5.15: Histogram showing the frequency distribution of the CMOS scores comparing the quality of (a) TM speech and NM speech, (b) TM speech and the speech synthesized using mapped LP coefficients and unmodified TM residual, $s_l(n)$, (c) $s_l(n)$ and the speech synthesized using mapped LP coefficients and modified TM residual, $\hat{s}_m(n)$, and (d) NM speech and $\hat{s}_m(n)$.

the NM speech ($\hat{R}_\eta(n)$) emphasized the instants in the vowel regions, and deemphasized the instants in the voiced consonant regions. Any undue emphasis/deemphasis caused by $\hat{R}_\eta(n)$, both in the voiced as well as unvoiced/silence regions was compensated by the gain parameter $\hat{G}_m(n)$. However, the results showed that a further deemphasis of the nasal segments is desired in order to obtain a better estimate of the envelope of the NM residual signal. The subjective evaluation showed that spectral mapping enhanced the perceptual quality of the TM speech. Residual modification resulted in a small improvement (compared to the spectral mapping) in the rating of the processed signal. It was shown that this technique, which involved spectral mapping and residual modification of voiced segments of the TM residual, improved the perceptual quality

of the TM speech. The advantage of this technique (spectral mapping and residual modification of voiced segments) is that it does not contribute any annoying artifact to the enhanced speech. The technique for residual modification needs to be improved further to achieve better results.

The TM speech is intelligible, and robust to the surrounding noise. This advantage of the TM speech is utilised for developing robust speech systems, as discussed in the following two chapters.

CHAPTER 6

RECOGNITION OF SOUND UNITS USING THROAT MICROPHONE SPEECH

There is a need to develop Automatic Speech Recognition (ASR) systems that perform reliably in adverse situations where normal microphones cannot be used. One example could be a small vocabulary, command-and-control ASR system in stealth operations. The acoustic analysis of the TM speech (refer Chapter 3) shows that there are distinguishing features in some of the sound units in the TM speech, such as the voiced stop consonants. It is possible to exploit the presence of such features to develop ASR systems using the TM speech. In this chapter, the feasibility of using the TM speech for speech recognition is studied. Previous studies (as explained in Section 2.2.3) used a multi-sensor approach, where features from the alternate speech sensors as well as the standard microphone were used to improve the performance of existing ASR systems. In contrast, this study evaluates the performance of ASR systems based on TM speech data. The performance of such a system is compared with that of an ASR system based on NM speech. This study would be useful to understand both the efficacy as well as the limitation of the TM speech for developing ASR systems.

This chapter is organised as follows. Section 6.1 briefly explains the choice of syllable as the recognition unit. Section 6.2 discusses the performance of an NM speech based syllable recognizer when tested with TM speech data as well as the estimated (using spectral mapping) NM speech data. Section 6.3 explains the performance of

the syllable recognizer developed using TM speech. The performance for syllable classification as well as for isolated utterances is discussed in this section.

6.1 SYLLABLE AS A SUBWORD RECOGNITION UNIT

This chapter focuses on Hidden Markov Model (HMM) based recognition of the 145 Consonant-Vowel (CV) units in an Indian language, Hindi. These 145 CV units (listed in Table 6.1) correspond to the combinations of the 29 consonants and 5 vowels in Hindi. These syllables are common to many of the Indian languages. Syllable is chosen as the sound unit for recognition because it is a subword unit motivated from the speech production and perception point of view [97, 98]. Other subword units like phonemes and triphones are unable to sufficiently model the coarticulation effects [99]. The syllables are inherently capable of capturing the coarticulation effects better than the phonemes and triphones. A syllable has a core vowel, and preceded or succeeded by one or more consonants. The structure of a syllable is $C^mVC^n,\ m, n \geq 0$. Recognition systems using syllables as basic units were shown to perform better than those using monophones as basic units for English [6] and Greek [100] languages. In [101–106], studies on modelling the syllabic units for the Indian languages were reported.

The number of syllabic units in a language is very large. The confusability among several CV units is high because of the similarities in their speech production mechanism. Since the syllabic units are larger in size compared to the phonemes, the frequency of occurrence of these units in continuous speech in usually low. Thus the number of training examples available to develop a model is limited. Consequently, developing robust models of these CV units for recognition is a difficult task.

Table 6.1: List of 145 CV units and their phonetic description features.

Manner Of Articulation (MOA)	Place Of Articulation (POA)	Vowel				
		/a/	/i/	/u/	/e/	/o/
UnVoiced UnAspirated (UVUA)	Velar	ka	ki	ku	ke	ko
	Palatal	ca	ci	cu	ce	co
	Alveolar	Ta	Ti	Tu	Te	To
	Dental	ta	ti	tu	te	to
	Bilabial	pa	pi	pu	pe	po
UnVoiced Aspirated (UVA)	Velar	kha	khi	khu	khe	kho
	Palatal	cha	chi	chu	che	cho
	Alveolar	Tha	Thi	Thu	The	Tho
	Dental	tha	thi	thu	the	tho
	Bilabial	pha	phi	phu	phe	pho
Voiced UnAspirated (VUA)	Velar	ga	gi	gu	ge	go
	Palatal	ja	ji	ju	je	jo
	Alveolar	Da	Di	Du	De	Do
	Dental	da	di	du	de	do
	Bilabial	ba	bi	bu	be	bo
Voiced Aspirated (VA)	Velar	gha	ghi	ghu	ghe	gho
	Palatal	jha	jhi	jhu	jhe	jho
	Alveolar	Dha	Dhi	Dhu	Dhe	Dho
	Dental	dha	dhi	dhu	dhe	dho
	Bilabial	bha	bhi	bhu	bhe	bho
Nasals	Dental	na	ni	nu	ne	no
	Bilabial	ma	mi	mu	me	mo
Semivowels	Palatal	ya	yi	yu	ye	yo
	Alveolar	ra	ri	ru	re	ro
	Dental	la	li	lu	le	lo
	Bilabial	va	vi	vu	ve	vo
Fricatives	Velar	ha	hi	hu	he	ho
	Alveolar	sha	shi	shu	she	sho
	Dental	sa	si	su	se	so

In the following section, the recognition performance of the standard HMM model based on the NM speech, when tested with the TM speech data as well as estimated NM speech data, is discussed.

6.2 RECOGNITION OF SYLLABIC UNITS OF TM SPEECH AND ESTIMATED NM SPEECH USING NM SPEECH BASED SYLLABLE RECOGNIZER

The performance of an existing isolated syllable recognizer, developed using the NM speech, for recognizing the CV utterances recorded using a throat microphone is studied. The recognizer is a left-to-right, no skip, HMM model with 5 emitting states developed for each of the 145 syllables. The recognition performance for the TM speech data is poor compared to that for the NM speech data (refer Table 6.2). This poor performance is due to the spectral mismatch in the TM and NM speech data for various sound units, as seen in Chapter 3. So, the HMM trained using the NM (wLPCCs) features performs poorly when tested with the TM (wLPCCs) features.

To improve the recognition performance, the estimated NM (wLPCCs) features (obtained by spectral mapping) are used for recognition. The performance of the recognizer improves with the estimated NM features, but is still not comparable to that of the NM features. In certain situations where normal microphones cannot be used, it is desirable to obtain a recognition performance that is close to the performance obtained when NM features are used. To realise such a recognition, a syllable recognizer is trained using the TM speech data.

6.3 SYLLABLE RECOGNIZER TRAINED USING TM SPEECH DATA

In this section, the performance of the TM speech based recognizer is studied in comparison to the NM speech based recognizer. The database for this study is comprised of isolated utterances of CV units of an Indian language, Hindi. Isolated utterances of CV units are chosen, instead of continuous speech data, to discount the effect of coar-

Table 6.2: Performance of the NM speech based syllable recognizer for the test data from NM, TM and estimated NM speech.

Test Features	N-best Recognition accuracy(%)				
	N=1	N=2	N=3	N=4	N=5
NM	39.03	55.07	63.83	69.03	73.72
TM	9.1	17.61	23.18	31.92	39.03
Estimated NM	18.7	29.41	37.16	46.83	51.28

ticulation and pronunciation variations. The speech data for 15 isolated utterances of each of the 145 CV units is collected from 5 male speakers. There are 75 utterances of each CV unit. The speech data is recorded using a sampling rate of 16 kHz, simultaneously using a throat microphone and a normal microphone. The recording is done in laboratory environment. A noise-free environment is chosen so as to study the feasibility of using the throat microphone for the speech recognition task. Among the 75 utterances for each CV unit, 11 examples from each speaker are used for training and 4 for testing. There are a total of $145 \times 11 \times 5 = 7975$ utterances for training and $145 \times 4 \times 5 = 2900$ utterances for testing, for the throat microphone and for the normal microphone. The speech data is processed to extract 13-dimensional wLPCC for every 15 msec frame, shifted by 5 msec.

6.3.1 Recognition of isolated utterances of CV units

As mentioned in Section 6.2, a standard left-to right, no skip HMM model with 5 emitting states is developed for each of the 145 units using the TM speech. A similar

Table 6.3: Performance of the isolated syllable recognizer based on the TM speech and the NM speech.

Microphone	N-best Recognition accuracy(%)				
	N=1	N=2	N=3	N=4	N=5
TM	31.84	45.03	52.15	57.15	61.77
NM	39.03	55.07	63.83	69.03	73.72

model is built, for comparison, using the simultaneously recorded NM speech. During testing, each utterance is tested in isolation and a syllable is hypothesised as one among the 145 CV units. Recognition of the CV units is tougher than the recognition of the E-set of English alphabet [104]. Though the recognition rate is not high due to the nature of the task, we observe that the performance of the recognizer based on the TM speech is comparable to that of the recognizer based on the normal microphone speech. The recognition performance of the TM speech based system is given in Table 6.3. The recognition performance of the NM speech based system, as in the first row of Table 6.2, is reproduced here for the purpose of comparison. The NM speech based system performs better than the TM speech based system. The difference in performance of the two systems is explained in the next section.

6.3.2 Classification of groups in different categories of CV units

A broad phoneme classification of the sound units is useful for some applications. For example, recognition of the broad phoneme classes may help in enhancing each sound class based on its acoustic-phonetic properties. Also, in some ASR applications it may be useful to group the syllables into broad classes as a preclassification step before

recognition of the syllable.

The CV units can be broadly classified into groups as shown in Table 6.1. A total of six systems are developed to categorise the speech data in each of the three categories, Place of Articulation (POA), Manner of Articulation (MOA) and vowel, for both the TM and the NM speech. Data pertaining to each group in a category is pooled together to develop an HMM for that group. The five POA groups are: velar, palatal, alveolar, dental and bilabial. The POA group recognizer would determine the group for a given test utterance. Likewise, the seven MOA groups are: unvoiced unaspirated (UVUA), unvoiced aspirated (UVA), voiced unaspirated (VUA), voiced aspirated (VA), nasals, semivowels and fricatives. The five vowel groups are: */a/, /e/, /i/, /o/* and */u/* in the syllable is also trained. The N-best recognition performance of each of the three group recognizers for the TM and NM speech based systems is given in Tables 6.4, 6.5 and 6.6. The performance of the group recognition systems for different groups in a category is given in Table 6.7. The confusion matrices for the POA, MOA and vowel groups for the TM and NM speech based systems are given in Tables 6.8 through 6.13.

The performance of the TM speech based group recognizers is similar to that of the NM speech based group recognizers for the case of N=1 (refer Tables 6.4 and 6.5). However, as N increases, the NM speech based system has a higher recognition rate for the POA category, while the TM speech based system has a higher recognition rate for the MOA category. This can be explained as follows. The acoustic cues for POA of the unvoiced sounds are not well picked up by the TM speech. This is because the articulation of these sounds involves a turbulence or a burst at some

location in the oral cavity, which the throat microphone cannot pick up as well as the normal microphone. Hence recognition of the POA group is poorer in the TM speech based system compared to the NM speech based system. In contrast, the MOA group recognition is better in the TM speech based system because the TM picks up the oral resonances behind the location of constriction (refer Section 3.3). These resonances, characterising the different MOA groups, are not picked up by the NM as it is placed in front of the constriction.

Within the POA category (refer Tables 6.8 and 6.9), the recognition of the velar group is better in the TM speech based system. This could be due to the proximity of the POA of these sounds to the location of the TM. The TM speech based recognition of the palatal and dental groups, and to a lesser extent the bilabial group is comparable to the NM speech based recognition. However, the recognition accuracy drops significantly for the alveolar group in the TM speech based system. This drop in recognition accuracy could probably be due to the burst and the succeeding transitional phase of the CV unit being better transmitted from the dental (dental sounds) and lip regions (bilabial sounds) to the throat region through the lower jaw, than from the alveolar ridge (alveolar sounds) (refer Fig. 1.1).

Within the MOA category, (refer Tables 6.10 and 6.11), the unaspirated (both unvoiced and voiced) groups are recognised much better in the TM speech based system than in the NM speech based system. Voiced stops are recognized better in the TM speech based system due to the presence of distinct formant-like structures during the closure phase of their articulation (refer Fig. 3.4). In contrast, all the voiced stops in the NM speech are characterised by the presence of a low frequency voice bar.

Table 6.4: Performance of the POA group recognizer based on the TM speech and the NM speech.

Microphone	N-best recognition accuracy(%)		
	N=1	N=2	N=3
TM	41.22	63.19	80.14
NM	41.49	66.11	83.06

Table 6.5: Performance of the MOA group recognizer based on the TM speech and the NM speech.

Microphone	N-best recognition accuracy(%)		
	N=1	N=2	N=3
TM	46.63	70.10	82.08
NM	46.04	66.84	78.89

Table 6.6: Performance of the vowel group recognizer based on the TM speech and the NM speech.

Microphone	N-best recognition accuracy(%)		
	N=1	N=2	N=3
TM	72.05	91.53	97.22
NM	87.88	99.41	99.76

Table 6.7: Performance of the group recognizers for different groups based on the TM speech and NM speech

POA group	Recognition accuracy (%)		MOA group	Recognition accuracy (%)		Vowel group	Recognition accuracy (%)	
	NM	TM		NM	TM		NM	TM
Alveolar	29.9	13.9	UVUA	49.2	64.4	/a/	96.5	83.8
Bilabial	52.7	41.4	UVA	47.4	28.9	/e/	88.6	76.8
Dental	33.6	50.6	VUA	48.4	66.7	/i/	90.8	61.5
Palatal	55.7	51.5	VA	36.1	24.3	/o/	76.3	80.8
Velar	38.8	50.3	Nasal	60.1	56.1	/u/	87.3	57.3
			Semi-vowel	39.9	38.2			
			Fricative	50.0	55.7			

Table 6.8: Confusion matrix (%) for the POA group classifier based on the TM speech.

Group	Alveolar	Bilabial	Dental	Palatal	Velar
Alveolar	14	9	28	23	26
Bilabial	3	41	15	14	27
Dental	8	9	51	12	20
Palatal	16	6	15	51	12
Velar	8	17	12	13	50

Table 6.9: Confusion matrix (%) for the POA group classifier based on the NM speech.

Group	Alveolar	Bilabial	Dental	Palatal	Velar
Alveolar	30	24	10	24	12
Bilabial	7	53	8	15	17
Dental	10	23	34	18	15
Palatal	27	6	8	56	3
Velar	10	22	7	22	39

However, the recognition of aspirated (both unvoiced and voiced) stops is poorer in the TM speech based system. They are confused with the unaspirated stop sounds. That is, unvoiced aspirated stops (e.g., */kha/*) are confused with unvoiced unaspirated stops (e.g., */ka/*), and unvoiced aspirated stops (e.g., */gha/*) are confused with voiced unaspirated stops (e.g., */ga/*). This is because the turbulence in the oral cavity during the aspiration phase following the release of the closure is not well captured by the TM (refer Fig. 3.3). The recognition performance for nasals and semivowels is comparable in the two systems. However the confusion between the nasals/semivowels and voiced unaspirated stops is higher in the TM speech based system than in the NM speech based system. This could be because the TM better captures the oral resonances behind the closure/constriction in these sounds. So the spectral features representing these sounds (i.e., nasals/semivowels vs. voiced unaspirated stops) in the TM speech are some times not distinguishable, especially if they have the same POA (e.g., */na/*

Table 6.10: Confusion matrix (%) for the MOA group classifier based on the TM speech.

Group	UVUA	UVA	VUA	VA	Nasal	Semi-vowel	Fricative
UVUA	64	7	13	1	4	1	10
UVA	31	29	16	0	7	3	14
VUA	7	0	70	10	9	0	4
VA	2	3	40	24	19	7	5
Nasal	12	0	21	3	56	7	1
Semi-vowel	15	1	30	2	13	38	1
Fricative	7	8	11	1	7	10	56

and */dha/*). In the case of fricatives, the performance of the TM speech based system is comparable to that of the NM speech based system. This shows that the missing higher frequencies in fricatives in the TM speech do not affect its recognition.

In the vowel category (refer Tables 6.12 and 6.13), generally the performance drops for most of the vowels in the TM speech based system, compared to the NM speech based system. In the NM speech based system the front vowels */i/* and */e/* are confused. Similarly, the back vowels */u/* and */o/* are confused. That is, the confusion is in the tongue height of the vowel for a given position of the tongue hump. In the TM speech based system, the vowels are confused not only based on the tongue height, but also based on the position of the tongue hump. For example, vowel */i/* is confused not only with */e/*, but also with */u/*. Similarly, vowel */u/* is confused with */o/* as well as

Table 6.11: Confusion matrix (%) for the MOA group classifier based on the NM speech.

Group	UVUA	UVA	VUA	VA	Nasal	Semi-vowel	Fricative
UVUA	49	18	3	4	5	10	11
UVA	11	47	1	6	6	6	23
VUA	8	1	48	14	7	18	4
VA	8	8	21	36	8	8	11
Nasal	15	2	8	5	60	7	3
Semi-vowel	13	5	16	12	9	40	5
Fricative	11	22	0	6	4	7	50

Table 6.12: Confusion matrix (%) for the vowel group classifier based on the TM speech.

Group	a	e	i	o	u
a	84	3	1	10	2
e	2	77	13	2	6
i	0	21	62	2	15
o	3	4	2	81	10
u	1	6	17	19	57

Table 6.13: Confusion matrix (%) for the vowel group classifier based on the NM speech.

Group	a	e	i	o	u
a	97	0	0	3	0
e	0	89	11	0	0
i	0	9	91	0	0
o	1	0	0	76	23
u	1	0	0	12	87

/i/. Among the vowels recognized by the TM speech based system, the recognition is better for the lower vowels (*/a/*, */e/* and */o/*) compared to the higher vowels (*/i/* and */u/*). In general, the performance of the TM speech based system in recognising the broad sound classes is comparable to the NM speech based system, with an improved performance in some of the classes which contain discriminating spectral features in the TM speech.

6.4 SUMMARY

In this chapter, a TM speech based syllable recognizer is developed, as it is desirable to have a speech recognition system in conditions where it may not be possible to use a normal microphone. When TM speech was tested against the existing NM speech based HMM syllable recognizer, the recognition rate was poor. Though the performance improved when tested with the estimated NM features, it is desirable to obtain a performance similar to that obtained when tested with NM features. As the

TM signal contains significant information about speech, HMM models were built and tested against TM speech. The performance of the TM speech based syllable recognizer was comparable to the performance of the NM speech based recognizer. For certain sounds, where the throat microphone captured additional spectral information, the recognition was higher.

The main advantage of the throat microphone speech is that it is robust to the ambient noise. In the next chapter, the TM speech has been used to develop robust speech systems in noisy conditions.

CHAPTER 7

ROBUST SPEECH SYSTEMS USING THROAT MICROPHONE SPEECH

7.1 INTRODUCTION

Robust speech systems are required for applications such as entry into high-security enclosures and access control in noisy environments that involve a reliable person identification. Telephone companies that handle calls from non-native speakers in noisy environments require reliable language identification. The performance of standard speech systems (using normal microphone speech data) degrades in noisy conditions. Approaches to improve the reliability of such systems have been reported (for example, [107, 108]). As mentioned in Section 1.3.2, in noisy conditions, a person's voice is less affected when recorded using a throat microphone than when recorded using a normal microphone. In this chapter, the throat microphone speech has been used to develop speech systems that are robust to noise. Two systems chosen to demonstrate this are: speaker recognition system and spoken Language IDentification (LID) system. Text-independent speaker recognition and LID systems are developed based on features extracted from the speech recorded using a throat microphone, and their performance is compared with that of systems based on normal microphone speech.

The differences in the dimensions and vibrations of the vocal folds among speakers play a major role in identifying the speaker [5, 79, 109]. Also, no two vocal tracts are of the same size and shape [109]. Hence, features that represent the excitation

source and vocal tract characteristics that are extracted from the normal microphone speech have been used for building speaker recognition systems [8, 9, 85, 110–115]. As the throat microphone is placed in proximity to the vocal folds, the speaker-specific excitation source characteristics may be captured by the throat microphone. Also, the TM speech contains significant information about the vocal tract characteristics, as seen in Chapter 3, which is expected to contain speaker-specific information.

While speaking, the shape of the vocal tract is associated with the phoneme that is articulated, and hence the language spoken. Even among languages with common phonemes, the pronunciation may vary. These pronunciation variations in phonemes among languages may be present in the vocal tract characteristics as well as the excitation source characteristics of the TM speech. In this chapter, the presence of speaker-specific and language-specific features in the vocal tract and excitation source characteristics of the TM speech is studied.

This Chapter is organized as follows. The speaker recognition studies are presented in Section 7.2. Section 7.3 describes the LID studies. The work on speaker recognition and LID is summarized in Section 7.4.

7.2 SPEAKER RECOGNITION USING TM SPEECH

Speaker recognition is the task of person identification using speech as the biometric feature [8, 110]. A person's voice, like other biometrics (finger prints, retinal patterns or genetic structure), cannot be forgotten or misplaced unlike the use of artifacts for identification by artificial means such as keys or memorized passwords [9] [111]. Hence, speaker recognition is more reliable than other artifacts for person identification. The

task of speaker recognition is to classify a spoken phrase as belonging to one among a set of reference speakers.

7.2.1 Speaker-specific features in TM speech

Parameters that represent the speaker-specific information in the vocal tract and excitation source characteristics are obtained using LP analysis. Typically, the vocal tract shape and size for different speakers could vary more at a finer level than at a gross level. So, parameters that can distinguish these finer variations may contain the speaker-specific information. The finer variations are dependent on the order of the LP analysis, which determines the number of peaks in the spectrum of an all-pole system. Each complex pole-pair accounts for one resonance peak, and the real poles account for the roll-off of the spectrum. A low (6th) order of LP analysis captures the gross features of the envelope of the speech spectrum. In contrast, a higher ($\geq 12^{th}$) order LP analysis captures the finer details along with the gross details of the envelope of the speech spectrum. Thus, speaker-specific features are captured by a higher order LP analysis [85]. The finer variations among speakers are observed more at higher frequencies, than at lower frequencies. In the context of vowels, it is mentioned in [79] that the exact positions of the higher formants vary a great deal from speaker to speaker. Though the formant positions are not uniformly determined for each speaker, they are certainly indicative of a person's voice. In the TM speech, since some of the higher frequencies have a low intensity, double differencing is done to emphasize the higher frequencies. A high order LP analysis will then provide speaker-specific information in the TM speech. The wLPCCs derived from higher order LP analysis are

used to represent the speaker-specific vocal tract characteristics in this study.

The movement of the vocal folds during vibration vary from one individual to another [114]. The variation could be in the extent of closure, and also in the manner and rate of closure [116]. For certain speakers, the vocal folds close completely, while for others the vocal folds never close completely. The vocal folds may close in a zipper-like fashion, or may close along the length of the vocal folds at approximately the same time. The rate of closure is faster in females compared to males. These speaker-specific variations in the vocal fold vibrations are assumed to be well captured by the throat microphone due to the proximity of its placement to the vocal folds.

The LP residual signal contains significant information about the excitation source characteristics of the speech signal. The voiced segments of the LP residual signal approximate the quasi-periodic vibrations of the vocal folds associated with the vibrations of the voiced sounds. As the variation among the speakers are associated with the vibrations of the vocal folds, the voiced segments of the LP residual segments are considered to contain speaker-specific information. In order to effectively capture the speaker-specific information from the voiced segments of the LP residual, it is necessary to minimize the presence of vocal tract characteristics in the LP residual signal. The LP analysis extracts the second order statistical features through the autocorrelation matrix. Since the second order statistics corresponds to the vocal tract system, the LP residual is assumed to preserve the source features in some nonlinear (higher order) relation among its samples. When the LP analysis order is low (say, 3), the LP spectrum picks up only the prominent peaks (refer Fig. 5.4). Hence the LP residual would have a large amount of vocal tract characteristics. If a very large LP order (say,

30) is used, then the LP residual may be affected by the spurious nulls in the spectrum of the inverse filter. An LP order of $12-14$ seems to be appropriate for a speech signal sampled at 8 kHz [116]. It is not clear as to which specific set of parameters needs to be extracted from the LP residual signal to represent the speaker-specific excitation source information. Hence, the LP residual signal itself is used to represent the excitation source information [116].

The feature vectors derived from the speech signal will form clusters in the feature space, and will have a certain distribution in the feature space. As the feature vectors represent the vocal tract and excitation source characteristics specific to a speaker, their distribution is expected to be different for different speakers.

The feature vectors representing the speech data have a complex distribution in the multi-dimensional feature space, and the surface representing this distribution may be highly nonlinear. The potential of artificial neural networks as nonlinear models is exploited to capture the characteristics of the vectors unique to a speaker from the given training data [84, 86]. Specifically, the autoassociative neural network models, which are feedforward neural networks that perform the task of autoassociation are used [117].

7.2.2 Speaker models using autoassociative neural networks

The task of autoassociation is to associate a given pattern with itself during training, and then recall the associated pattern when an approximate version of the same pattern is given during testing [84]. A five layer Autoassociative Neural Network (AANN), comprising of three nonlinear hidden layers, is capable of modelling any arbitrary

distribution of the feature vectors [112, 117, 118]. The number of units in the input and output layers is equal to the dimension of the input feature vector. The middle hidden layer, which comprises of lesser number of units than the input/output layer, is a dimension compressing layer. The training error surface relates to the distribution of the given feature vectors [118]. As the 5-layer AANN is capable of modelling any arbitrary distribution, it is able to capture the distribution even when raw data like the LP residual is given [116]. Typical structure of a five layer AANN used in this study is shown in Figure 7.1.

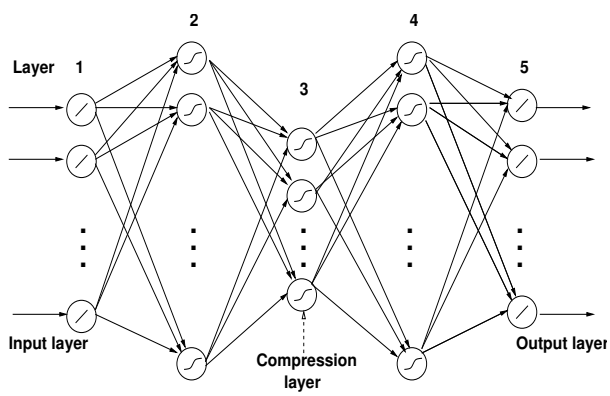

Figure 7.1: Five layer AANN model.

7.2.3 Speaker recognition - experimental study

7.2.3.1 Database for the study

The database for this study comprises of recordings done in the laboratory under clean and noisy conditions. Noisy environment is simulated using radio statics. Speech from volunteers is acquired simultaneously using the throat and normal microphones. Text-independent speech is used in this study. Two minutes of speech data obtained from each of the 40 speakers is used to train a speaker model. Each test utterance is of

20 seconds duration. The recordings for training and testing the speaker models are carried out in separate sessions. The 240 test utterances obtained from the 40 speakers under clean and noisy conditions are tested against each of the 40 speaker models.

7.2.3.2 Training the speaker models

Two separate models are built for each speaker, one that captures the distribution of the vocal tract system characteristics (system model) specific to the speaker, and the other that captures the distribution of the excitation source characteristics (source model) of the speaker. The vocal tract characteristics are represented by a 19-dimensional vector of wLPCCs, and the excitation source characteristics are represented by the LP residual.

The structure of the neural network used to model the system features is *19L 38N 4N 38N 19L*, where L refers to a linear unit, N to a nonlinear unit, and the numbers represent the number of nodes in a layer. The 19-dimensional vectors of wLPCCs obtained for each speaker are given to the AANN in a randomized fashion. Each AANN is trained for 200 epochs.

The LP residual is down-sampled to 4 kHz sampling frequency to emphasize only those regions with a high signal-to-noise ratio. Blocks of 20 samples (5 msec) of the normalized LP residual, with a shift of one sample, are applied in succession. The structure of the AANN used is *20L 40N 10N 40N 20L*. The model is trained for 200 epochs.

7.2.3.3 Testing the speaker models

Test utterances from clean and noisy environments, each of 20 seconds duration, are used to test the source and system based speaker models trained using the clean speech. The noisy test utterances are used to test the speaker models trained using the noisy speech. Both the source and system features are extracted using a 12th order LP analysis. The 19 dimensional vectors of wLPCCs and blocks of 20 samples of the LP residual shifted by one sample form the input to the system and source models, respectively [112]. The deviation of the output of each model from its input is used to compute the squared error E_i for the ith frame or block. This error is used to compute the confidence score for that frame or block, which is used as a performance measure for the speaker recognition system. The confidence score C_i of the ith frame or block is expressed as $C_i = \exp(-\lambda E_i)$ where the constant λ is set to 1 in this study. This confidence value is higher if the error is lower, when the frame or block of the test utterance of a speaker matches with the corresponding model. When the frame or block of the test utterance does not match with the corresponding model, the error is high, and this lowers the confidence score of that frame or block. A test utterance is given as input to each speaker model to obtain the average confidence score C, which is expressed as $C = \frac{1}{N}\sum_{i=1}^{N} C_i$, where N is the total number of frames/blocks. The average confidence scores of the test utterance against all the models are compared, and the model which gives the highest value of C corresponds to the hypothesized speaker.

7.2.3.4 Performance evaluation

Performance of the speaker recognition systems based on the system and source features derived from the clean speech obtained from throat and normal microphones is given in Table 7.1. The performance is evaluated in terms of percentage of the number of test utterances accepted out of the total test utterances used for this study. It can be seen that the performance of the speaker recognition system using the TM speech is similar to that using the NM speech for both the system features as well as the source features. As the scores obtained for both the system and source feature based models are from independent sources of evidence, the two scores can be combined. The combination logic is addition of the two scores. The block diagram of the proposed speaker recognition system using the combined evidence is shown in Fig. 7.2. The speaker recognition system based on the combined scores performs relatively better in the case of TM speech than for the NM speech. This may be due to the presence of significant speaker-specific information in the excitation source characteristics derived from the TM speech.

Table 7.2 shows the performance of the speaker recognition systems using the speaker models trained and tested with the noisy utterances (to ensure that the speaker characteristics are the same during training and testing). The performance of the TM speech based speaker recognition systems is similar under the clean and noisy conditions (refer Table 7.2). But, the performance of the NM speech based speaker recognition systems is poor in noisy conditions. This is due to degradation in the NM speech. The performance of NM speech based systems in noisy conditions can

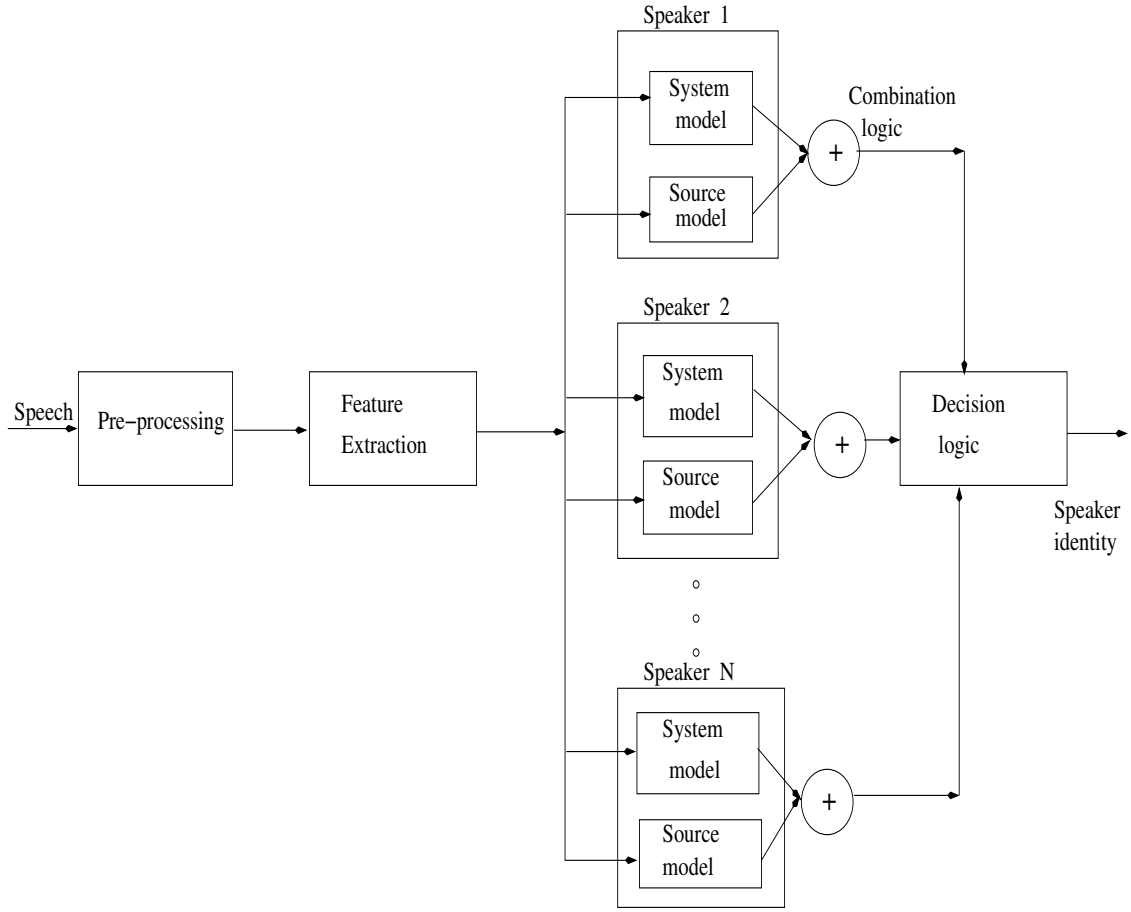

Figure 7.2: Block diagram of the speaker recognition system using the combined evidence.

be improved by enhancing the noisy NM speech prior to recognition. The improvement in performance will depend on the type of noise and the enhancement technique used. This study shows that the TM speech can be used for recognition without any preprocessing in noisy conditions.

7.3 LANGUAGE IDENTIFICATION USING TM SPEECH

Languages differ in the inventory of phonemes used to produce words, the frequency of occurrence of these units, and the order in which they occur in words [10]. Languages also differ in the duration of phonemes, speech rate, intonation, phonotactics (rules

Table 7.1: Performance (%) of the speaker recognition systems based on the source and system features obtained from simultaneously recorded speech signals using throat and normal microphones. Models are trained and tested using speech in clean environment.

Microphone	System features	Source features	Combined score
TM	84.3	73.0	94.3
NM	84.3	70.0	88.6

Table 7.2: Performance (%) of the speaker recognition systems based on the source and system features obtained from simultaneously recorded speech signals using throat and normal microphones. Models are trained and tested using speech in noisy environment.

Microphone	System features	Source features	Combined score
TM	83.3	75.0	93.3
NM	19.42	25.0	25.0

that govern the combination of different phonemes in a language), and vocabulary. The information related to these differences among the languages is present in the speech signal.

Automatic language identification (LID) is the task of identifying the language from a spoken utterance. Language identification systems have many applications like helping telephone companies in handling foreign language calls, serving as a front-end device for a multi-lingual speech recognizer, and a multi-language speech translation systems among others [11]. Various approaches have been proposed in the literature to discriminate between languages based on the differences present in the speech signal [10, 119–121]. Generally, language identification studies use a database comprising of uncorrupted NM speech or telephone speech [122]. In noisy environments, the LID systems may not perform well, as the input speech to the system is corrupted by noise. A few approaches to improve the performance of LID systems in noisy conditions have been proposed [108, 123]. In this section, the clean speech from the throat microphone is used to obtain a reliable LID system in noisy conditions. The presence of language-specific features in the vocal tract and excitation source characteristics of the TM speech are explored.

7.3.1 Language-specific features

The inventory of phonemes is not unique for a language. There is considerable overlap in the phonemes and syllables of various languages. But there are differences in the way the same phoneme or syllable is pronounced in different languages. This variation in pronunciation between languages may be reflected in the slight variation in

the articulatory configurations of the vocal tract associated with these phonemes (or syllables). Such variation in the vocal tract configurations can be represented using the short-term spectrum. The excitation source corresponding to the articulation of the sound units is also expected to contain information related to the language [124]. The presence of language-specific features in the vocal tract and excitation source characteristics of the speech are explored in this study.

The features extracted from the speech signal contain information pertaining not only to the language, but also the speaker and the sounds spoken. It is necessary to extract features that are more language-specific, rather than speaker-specific. This is achieved as follows. A higher order (≥ 12) LP analysis better captures the speaker-specific characteristics, as seen in Section 7.2.1. A lower order (say, 8) captures less of the speaker characteristics, and more of the sounds spoken and the language. To better capture the language characteristics, the feature vectors are derived from a concatenation of speech from an equal number of native male and female speakers of the language. This concatenation helps to capture the variability in speakers of a language. Also, the legal sound units of the language are expected to be captured. The distribution of the feature vectors is expected to be different for different languages. However, the distribution may be complex due to the similarity of the phonemes across languages. The AANN model is used to capture this complex distribution.

7.3.2 LID - experimental study

The language identification study compares the performance of the LID sytems based on the TM speech and the NM speech both in the clean and noisy conditions.

7.3.2.1 Multi-language speech corpus

The corpus comprises of recordings from 80 native speakers in four Indian languages, namely, Hindi, Kannada, Telugu and Tamil. All the languages belong to the same family of languages and share a common set of phonemes. The confusability among them is likely to be high. The recordings are carried out in the laboratory under clean and simulated noisy conditions. The noisy conditions are simulated using radio static. Speech from the volunteers is collected simultaneously using a normal microphone and a throat microphone. Text-independent speech is used in the study. The speech is sampled at 8 kHz. Features are extracted using an 8th order LP analysis performed on overlapping Hamming windowed speech frames of 20 msec duration taken with a frame shift of 5 msec duration.

7.3.2.2 Language identification using system features

In this study, the AANN model is used to capture the distribution of language-specific spectral and source feature vectors. The n feature vectors given as input to train each language model are derived from a concatenation of speech obtained from 6 speakers, 3 male and 3 female, such that $n/6$ feature vectors are derived from each speaker's data. Around 30 seconds of speech data from each speaker is used for training. The structure of the neural network used to capture the distribution of vocal tract system feature vectors in the feature space is *12L 38N 4N 38N 12L*. The 12-dimension system features (wLPCCs) obtained for each speaker are given to the AANN in a randomized fashion. Each AANN is trained for 200 epochs. Two separate models for each language are obtained by training two AANN models using WLPCCs derived from the clean and

noisy speech. If a system is to perform accurately under the noisy conditions, it needs to be trained on speech recorded under these conditions. Hence, noisy speech models are used. During testing, features derived from utterances of 20 secs duration, are used to compute confidence scores (as explained in Section 7.2.3.3) against models of each language and to identify the language of that utterance. The test speech utterances are different from those used for training.

7.3.2.3 Language identification using source features

The structure of the AANN used is *20L 40N 10N 40N 20L*. The LP residual, that represents the excitation source features is down-sampled to 4 kHz sampling frequency to emphasize only those regions with high signal-to-noise ratio are used. Blocks of 20 samples (5 msec) of the normalized LP residual, with a shift of one sample, are applied in succession to train the model. A model is trained for 500 epochs. Two models for each language are obtained using the LP residual derived from the clean and noisy speech.

During testing, the error between the output of the AANN and the input is used to compute the confidence score C_i for the ith frame. The average confidence scores of the test utterance against all the models are compared to evaluate the performance of the LID systems. The performance is evaluated in terms of percentage of the number of test utterances accepted out of the total number of test utterances used for this study.

7.3.2.4 Results and discussion

The performance of the LID systems based on the system and source features derived from the clean speech obtained from throat and normal microphones is given in Tables 7.3 and 7.4 respectively. It is seen that the performance of the LID system using the TM speech is similar to that using the NM speech for both the system features as well as the source features. As the scores obtained for both the system and source feature-based models form independent sources of evidence, the two scores are combined (by addition of the two scores) to improve the performance of the LID system. The block diagram depicting the LID system is shown in Figure. 7.3.

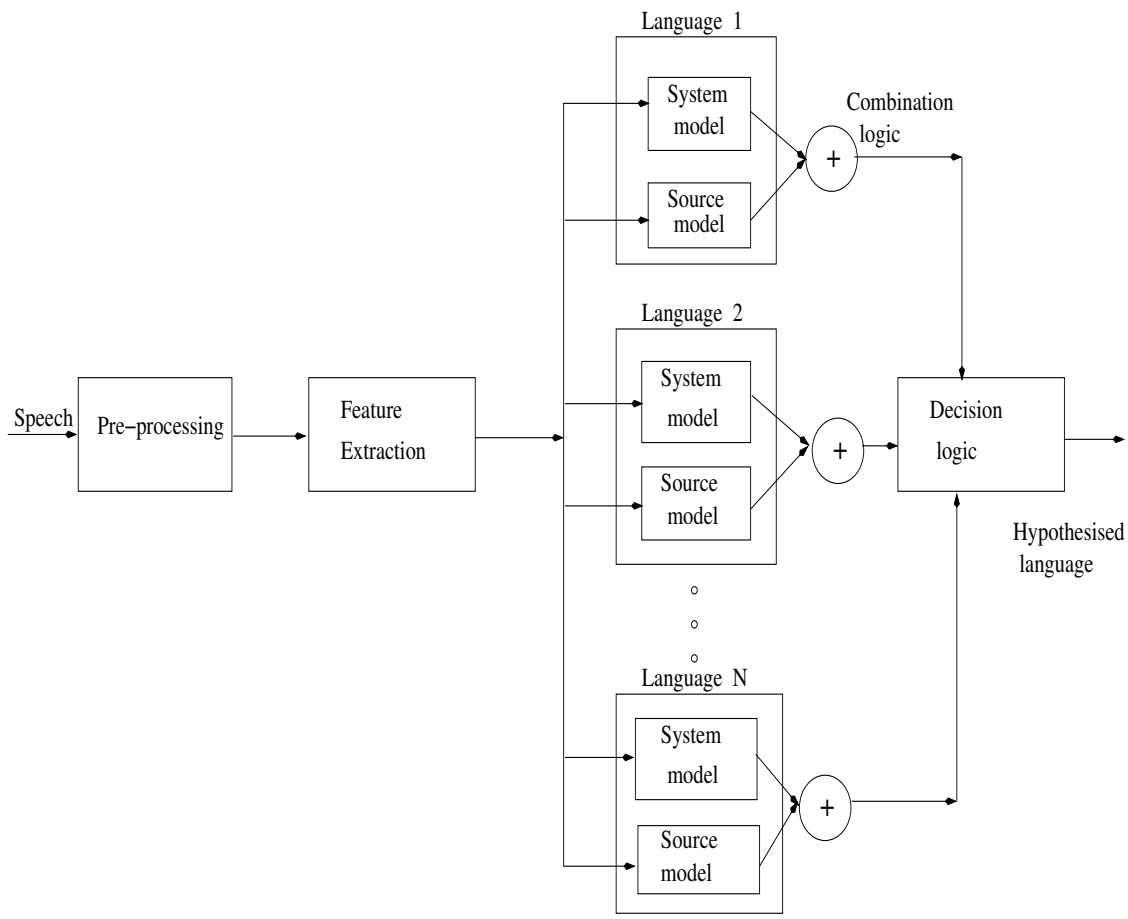

Figure 7.3: Block diagram of the language identification system using combined evidence.

Tables 7.5 and 7.6 show the performance of the LID systems, where the language models trained and tested using noisy speech. The performance of the TM speech based LID systems are similar in clean and noisy conditions (refer Table 7.3 and 7.5). But, the performance of the NM speech based LID systems degrades. This is due to the presence of significant noise levels in the speech. As mentioned in Section 7.2.3.4, the performance of the NM speech based LID system in noisy conditions can be improved by enhancing the NM speech prior to identification. The increase in performance will depend on the level of degradation and the enhancement technique used. But, this study shows that the TM speech based LID system does not require such preprocessing.

The studies show that the vocal tract system and excitation source features extracted from the TM speech contains significant information to distinguish languages. The performance of the language identification systems using TM and NM speech signals, when the signals are recorded under noise-free conditions is almost the same. However, under noisy conditions, the performance of the LID system using the TM speech is similar to that under noise-free conditions, while that using the unprocessed NM speech degrades.

7.4 SUMMARY

From the studies conducted, it was seen that the vocal tract system and excitation source characteristics derived from the TM speech indeed contains significant information about the speaker as well as the language. The performance of the speaker recognition and LID systems based on the TM and NM speech was almost similar

Table 7.3: Performance (%) of the language identification systems based on source and system features obtained from the speech recorded using a throat microphone in a clean environment.

Language	System features	Source features	Combined score
Hindi	90	69	92
Tamil	95	74	95
Telugu	92	70	96
Kannada	95	72.5	95

Table 7.4: Performance (%) of the language identification systems based on source and system features obtained from the speech recorded using a normal microphone in a clean environment.

Language	System features	Source features	Combined score
Hindi	87	71	91.5
Tamil	90	67	93
Telugu	96	77.5	100
Kannada	100	72	100

Table 7.5: Performance (%) of the language identification systems based on source and system features obtained from the speech recorded using a throat microphone under noisy conditions.

Language	System features	Source features	Combined score
Hindi	89	67.5	91
Tamil	90	76.5	94
Telugu	85	69	93.5
Kannada	92	75	92.5

Table 7.6: Performance (%) of the language identification systems based on source and system features obtained from the speech recorded using a normal microphone under noisy conditions.

Language	System features	Source features	Combined score
Hindi	57	56.5	57.5
Tamil	61	59	61
Telugu	45	52.5	51
Kannada	66.5	62	71.5

when the speech signals were recorded under noise-free conditions. The performance of the TM speech based speaker recognition system improved when the evidences from the system and source models were combined. This could be due to the presence of significant speaker-specific characteristics in the TM speech residual signal. Under noisy conditions, it was seen that the performance of the TM speech based systems was similar to the performance under noise-free conditions. However, the performance of the NM speech based systems reduced due to degradation in the NM speech. The performance of the NM speech based system would improve in noisy conditions if the noisy NM speech is enhanced prior to recognition. This study showed that the TM speech can be used for speech applications in noisy conditions without any need for enhancing the speech.

The techniques used to improve the perceptual quality of the TM speech can be extended for other applications, as discussed in the following chapter.

CHAPTER 8

BANDWIDTH EXTENSION OF TELEPHONE SPEECH AND LOUDNESS ENHANCEMENT

8.1 INTRODUCTION

In Chapters 4 and 5, techniques were proposed to improve the perceptual quality of the throat microphone speech. These techniques, which included mapping the spectral features of TM and NM speech, and modifying the LP residual of the voiced segments of the TM speech, can be extended for other applications. Speech from an analog telephone channel has a limited bandwidth (300 Hz to 3400 Hz). There is a need to improve its perceptual quality, so as to provide the subscribers of the analog telephone system a high quality speech, similar to the normal wideband speech. The spectral mapping approach and the residual modification of voiced segments can be extended for this task. Increasing the loudness of soft voices would be a practical consideration for power-limiting devices such as cell phones or hearing aids with high audio output requirements. The residual modification technique can be extended to enhance the loudness of soft voices. This chapter is organized as follows. Section 8.2 explains the approach for improving the perceptual quality of the narrowband telephone speech. The loudness enhancement of soft voices is discussed in Section 8.3. The work is summarized in Section 8.4.

8.2 BANDWIDTH EXTENSION OF NARROWBAND TELEPHONE SPEECH

Speech has perceptually significant energy in the 100-8000 Hz range. However, when this signal is passed through the analog telephone channel, it gets band-limited between 300-3400 Hz (refer Section 1.4). Digital networks such as Integrated Service Digital Network (ISDN) and Global System for Mobile communication (GSM) [4] are able to transmit higher quality speech, since the signal components below 300 Hz as well as components between 3400 Hz and 4000 Hz can also be transmitted. However, this is true if the entire call (in terms of routing) remains in these networks. When the signal leaves the digital network into an analog telephone network, the speech signal is once again band-limited.

Improvement in the perceptual quality of the telephone speech is done at the receiver side, without modifying the existing telephone network. The basic idea of enhancement is to estimate the speech signal components above 3400 Hz, and complement the signal in the idle frequency bands with this estimate. Most of the current bandwidth extension schemes use the source-filter model of speech production. Here, the bandwidth extension is divided into two separate tasks. One task is the recovery of the wideband spectral envelope, and the other is the regeneration of the wideband residual signal, as shown in Fig. 8.1. The wideband residual is then passed through the wideband LP synthesis filter to produce the wideband speech signal. The various approaches to recover the wideband spectral envelope and regeneration of the wideband residual signal are detailed in Section 2.3.

In this work too, the source-filter model is used. The recovery of the wideband

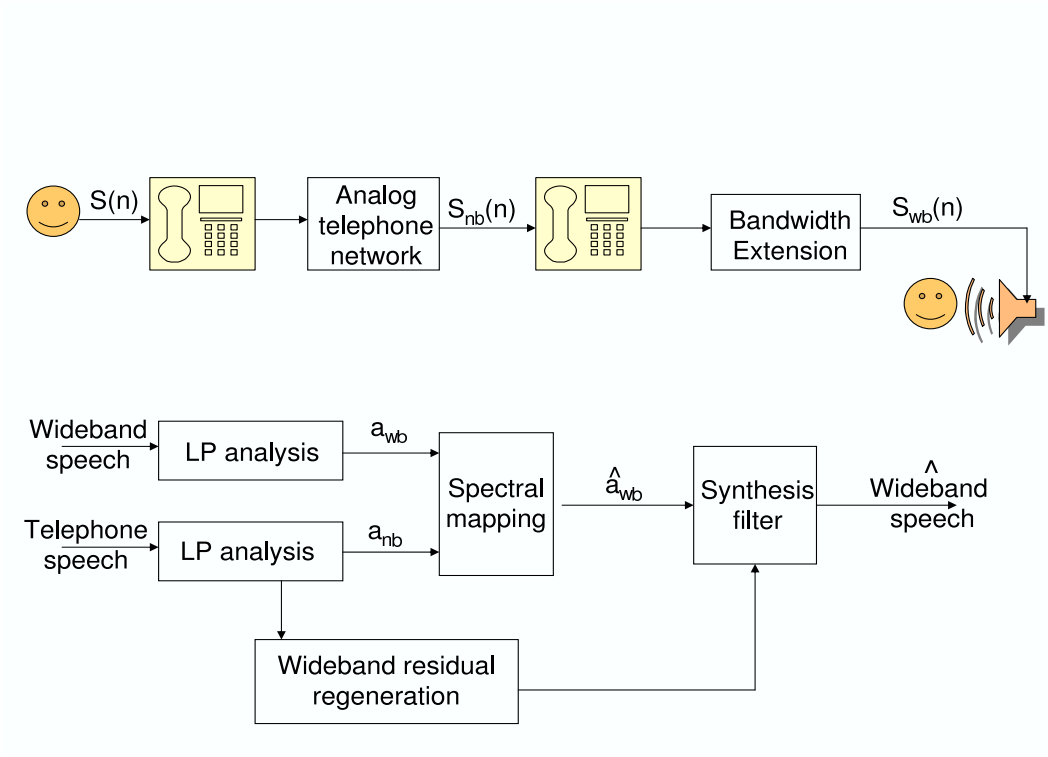

Figure 8.1: Schematic representation of the approach for bandwidth extension of telephone speech. Here, S refers to the wideband speech, S_{nb} refers to the narrowband speech, and S_{wb} refers to the estimated wideband speech.

spectral envelope involves a mapping of the band-limited spectral features onto the wideband (0-8000 Hz) spectral features. The mapping exploits the nonlinear mapping property of MLFFNN, as used to map the TM and NM spectra (refer Chapter 4). To regenerate the wideband residual, the narrowband residual is first preprocessed so as to reduce the distortion that would be otherwise present in the reconstructed speech. The distortion is caused due to the large linear prediction error of the telephone speech. The preprocessing involves modifying the voiced segments of the narrowband residual to reduce the distortion.

8.2.1 Recovery of wideband spectral envelope - mapping narrowband spectra to wideband spectra

The data for this study comprises of speech simultaneously recorded from a normal microphone at the transmitting end, and a telephone at the receiving end. A speaker speaks simultaneously into the microphone and the telephone. The data that is recorded by the normal microphone forms the wideband data. The data that is collected at the receiver end forms the narrowband data. Speech data for a duration of 5 minutes from each speaker is used to obtain the speaker-dependent models. The remaining sentences of the speaker are used to test the model. All the recorded speech signals are sampled at 16 kHz.

The LP analysis is performed on both the narrowband and wideband signals. The LP analysis window is 20 msec long, with a 10 msec overlap between successive windows. The LP order for the analysis of the narrowband data is 12, and the LP order for the analysis of the wideband data is 16. The LP coefficients are used to derive 20 dimensional wLPCCs. The wLPCCs derived from the narrowband telephone speech are mapped onto the wLPCCs derived from the wideband speech. The mapping is performed using the procedure described in Chapter 4. To accelerate the training process of the network, each training pattern is preprocessed so that the input-output pattern pairs will have zero mean and unity standard deviation [86]. They are then scaled so that they always fall within the range $[-1, 1]$. This ensures that the different synaptic weights of the network learn at approximately the same speed. These zero mean, scaled wLPCCs derived from the band-limited signal and the wideband signal form the input-output training pairs, respectively, for the mapping network.

The structure of the MLFFNN used in this study is $20L30N30N20L$. The batch mode of training is used here. The weight updation is done using the conjugate gradient method, as explained in Section 4.3.2.2. The network has been trained for 50 epochs.

During testing, wLPCCs of the test narrowband data are given to the mapping network. The network produces an output which are the estimated wLPCCs of the corresponding wideband data. The estimated wideband LP coefficients derived from these wLPCCs (as mentioned in Section 4.3.1) are used to construct the wideband LP synthesis filter.

Fig. 8.2 shows that the estimated LP spectrum is a close approximate of the LP spectrum of the wideband speech. This shows that the mapping network has efficiently captured the nonlinear mapping in the frequency domain between the narrowband and wideband speech.

8.2.2 Regeneration of wideband LP residual

Inverse filtering of the narrowband telephone speech emphasizes the higher frequencies in the LP residual signal. This is because the spectral roll-off is steep in the telephone speech compared to the wideband speech, as some of the higher frequencies are missing in the telephone speech. If this residual signal is used directly for regenerating the wideband residual, the reconstructed speech would sound slightly noisy. So it is necessary to modify the LP residual of the telephone speech to reduce the noise due to the emphasis of the higher frequencies, prior to regeneration of the wideband residual.

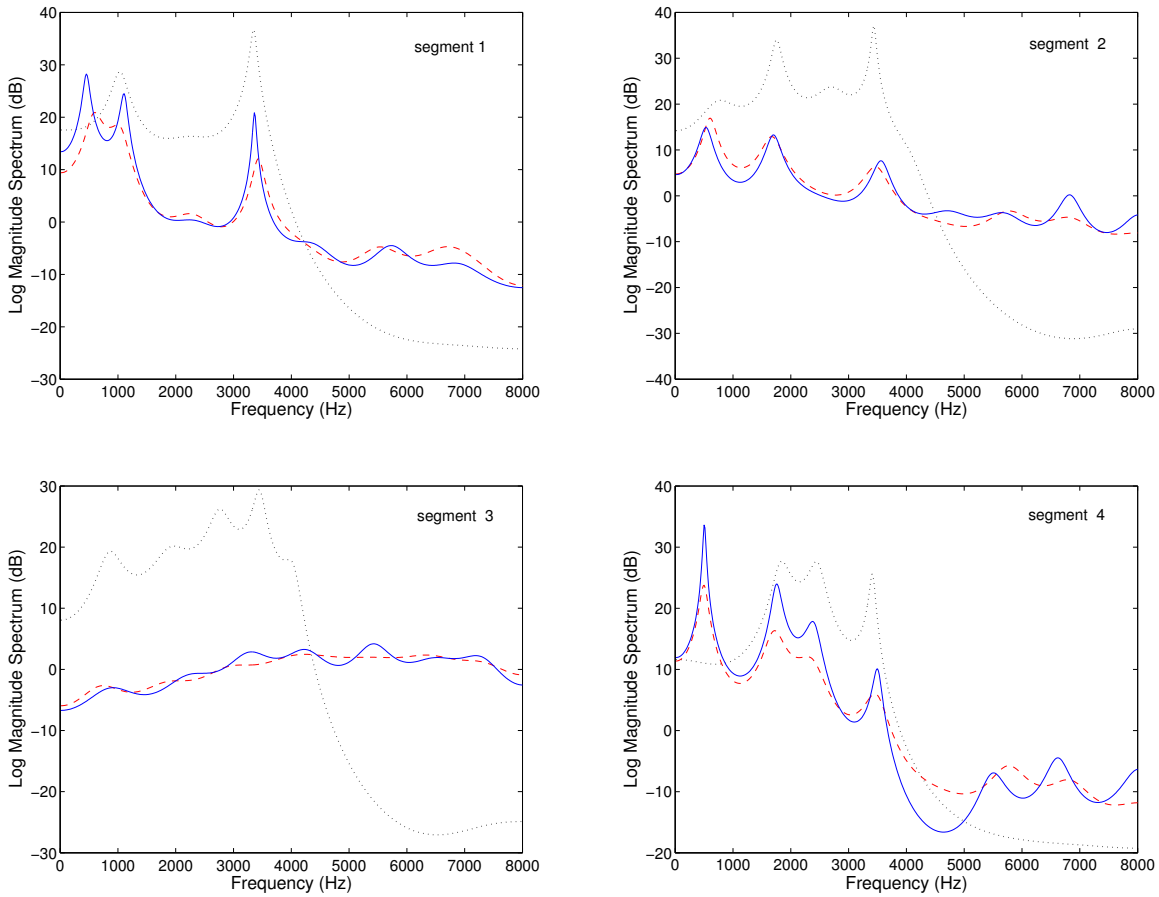

Figure 8.2: The estimated wideband LP spectra (dashed line), and the LP spectra of the narrowband telephone speech (dotted line) and wideband normal speech (solid line) for four segments of speech.

8.2.2.1 Deemphasis of noisy regions in LP residual of telephone speech

In the LP residual signal, the closed glottis portions correspond to high SNR regions, and the open glottis portions correspond to low SNR regions within each glottal cycle. The emphasis of the high frequency content in the narrowband residual signal affects the open glottis regions more than the closed glottis regions. This can be seen in Fig. 8.3. While in the wideband signal the open glottis regions are deemphasized relative to the closed glottis regions, in the narrowband telephone signal such a relative

deemphasis of the open glottis regions is not seen. In order to compensate for this, the signal in the open glottis regions needs to be deemphasized relative to the closed glottis regions. Human perception is based on capturing some features from the high SNR regions of speech, and extrapolating the features of the low SNR regions [5]. By performing a relative emphasis of the high SNR regions over the low SNR regions of the residual signal, the perceptual cues present in the speech could be emphasized. As correlation between the samples of the residual signal is low, this modification would not produce noticeable distortion in the reconstructed speech. This modification of the narrowband residual needs to performed on all voiced segments, irrespective of the type of sound, whereas the modification of the residual of the voiced segments of the TM speech was based on the broad phoneme category to which the voiced segment belonged (refer Section 5.2).

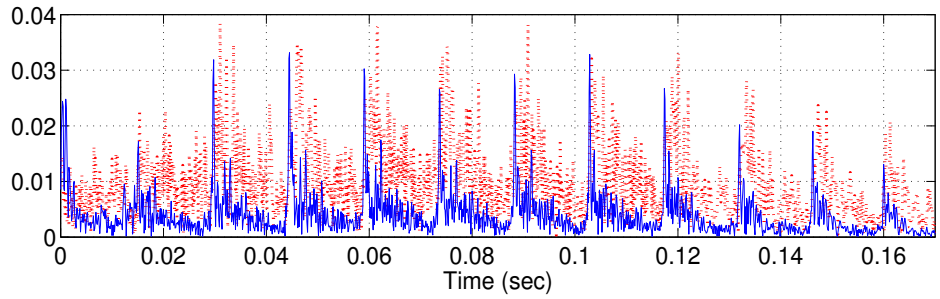

Figure 8.3: The Hilbert Envelope of the residual signal of the telephone speech (dashed line) and the wideband speech (solid line) of a speech utterance. The open glottis regions (2-3 ms before the instants) appear noisy in the telephone signal, compared to the wideband signal.

In order to perform a deemphasis of the open glottis regions relative to the closed glottis regions, the instants of glottal closure have to be identified. These instants of glottal closure are not identified directly from the LP residual, though the LP residual

contains information pertaining to the excitation (refer Section 5.5). The Hilbert envelope of the LP residual gives a better estimate of the instants of glottal closure.

Once the instants have been located, a region of 2 to 3 msec following an instant is chosen as the closed glottis region. A region of 2 to 3 ms prior to the instant is chosen as the open glottis region. A weight function is used to deemphasize the excitation in the open glottis regions relative to the closed glottis regions as explained below.

The Hilbert envelope of the LP residual is smoothed using a window which is selected to taper down the excitation in the open glottis region, and raise the excitation around the glottal closure (GC) instant. This window is given by

$$w(n) = 0.54 - 0.46\cos\frac{2\pi}{M}(n + M/2), \quad 0 \leq n \leq M/2, \tag{8.1}$$

where $M/2$ is size of the window. Here, $w(n)$ corresponds to the second half of a Hamming window. The smoothing operation produces weights which depend on the strengths of the peaks in the Hilbert envelope signal. In order to normalize the weights, the smoothed signal is scaled by the running mean of the signal to obtain the normalized signal $C(n)$. The signal $C(n)$ is mapped using a nonlinear sigmoidal mapping function to obtain a weight function $W(n)$ given by [125]

$$W(n) = \left(\frac{1 - \zeta_{min}}{2}\right)\tanh(\alpha\pi(C(n) - \zeta_0)) + \left(\frac{1 + \zeta_{min}}{2}\right). \tag{8.2}$$

The parameter ζ_{min} is used to reduce the excitation signal in the open glottis regions. The parameter α is used to taper the weight function smoothly from the closed glottis region to the open phase region to avoid distortion. The parameter ζ_0 is

the threshold value that determines the weight values to be emphasized or suppressed. The weight values are adjusted such that the closed glottis segment is given a higher weightage, while the excitation in the open phase region is reduced (refer Fig. 8.4). The LP residual of the narrowband telephone speech is multiplied with this weight function to obtain the modified LP residual.

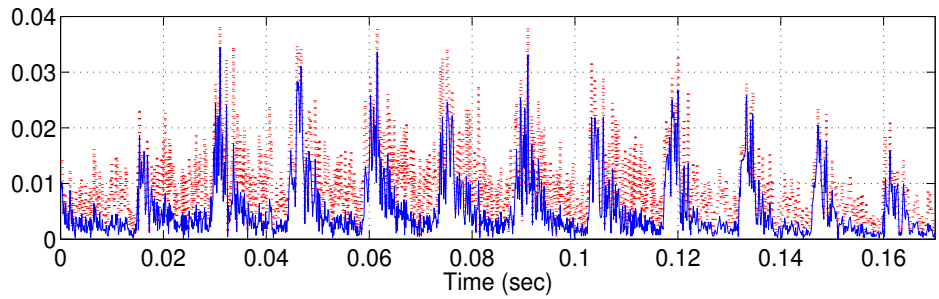

Figure 8.4: The Hilbert Envelope of the residual signal of the telephone speech (dashed line) and the modified narrowband residual (solid line) of a speech utterance. The open glottis regions are deemphasized in the modified signal, compared to the narrowband signal.

8.2.2.2 Regeneration of wideband residual using spectral folding

The modified narrowband LP residual is used to regenerate the wideband LP residual. This regeneration is based on duplication of the baseband spectrum [61]. The LP residual has a flat line spectrum at multiples of the fundamental frequency. Such a spectral structure is periodic and repetitive. The high frequency structure is the same as the low frequencies. The spectrum of unvoiced excitation is however continuous and has a random spectrum with a flat envelope. The details of the unvoiced spectrum are not as perceptually important as the details of the voiced spectrum. Therefore the unvoiced spectrum can be considered repetitive also. To perform spectral folding, zero - valued samples are inserted between samples of the narrowband modified LP

residual. This process is merely that of upsampling which produces spectral folding. This is a simple technique. Other sophisticated regeneration methods could also be used.

The regenerated wideband residual is passed through the LP synthesis filter to obtain the wideband speech signal. This is passed through a high pass filter and then added to the band-limited signal in order to preserve the frequency components in the lower frequency bands to produce the wideband signal.

8.2.2.3 Experimental results

The spectrograms for the original wideband speech, the band-limited speech and the enhanced speech for a segment of speech are shown in Fig. 8.5. It can seen the spectral information present in the higher frequency range (above 4000 Hz as in fricatives) in the original wideband speech is also present in the enhanced speech. The lower frequency information is also preserved in the enhanced signal.

8.2.2.4 Subjective evaluation

The perceptual quality of the enhanced speech is assessed using the CMOS test. The CMOS test is explained in Section 5.7. The results of the evaluation are shown in Fig. 8.6 in the form of histograms. The listeners rated the wideband speech with a CMOS of 1.8 (between slightly better and better) over the narrowband telephone speech. The enhanced speech, obtained using the approach described in the previous sections, is rated with a CMOS of 0.78 (between 'about the same' and 'slightly better') over the telephone speech. This shows that the technique used to extend the bandwidth of the narrowband telephone speech provides an improvement in the perceptual quality of the

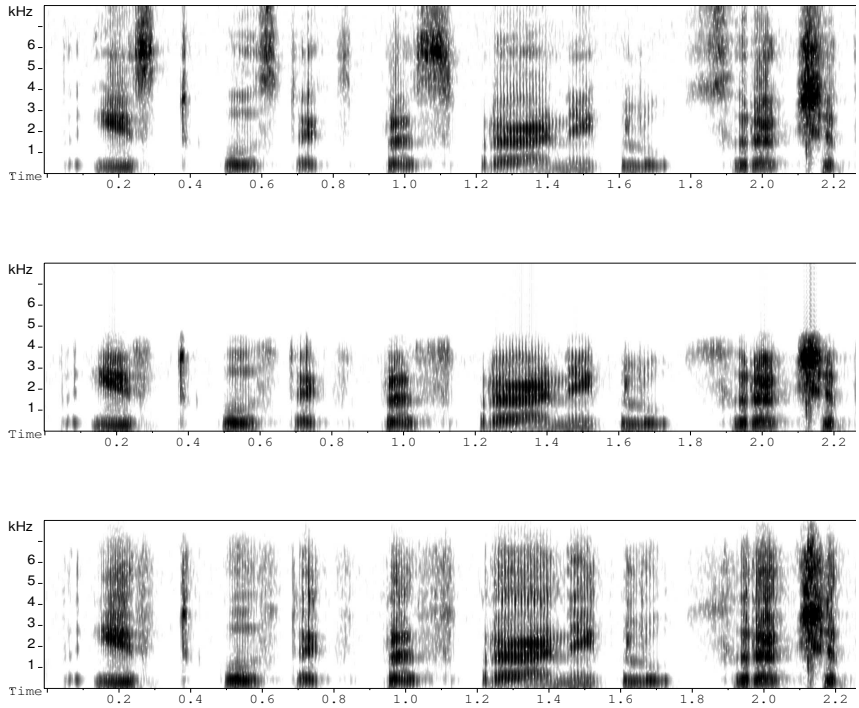

Figure 8.5: Spectrograms of the (a) wideband speech signal, (b) band-limited speech signal, and (c) the estimated wideband speech signal for a speech segment.

speech. The technique used for bandwidth extension does not contribute any annoying artifact to the enhanced speech, as in the enhancement of the throat microphone speech.

8.3 LOUDNESS ENHANCEMENT OF SOFT VOICES

Enhancing the loudness of soft voices is a topic of interest for the manufacturers of cell phones and other devices with high audio output such as hearing aids. Design of devices with a low cost and limited consumption of power is a major issue. Limiting the consumption of power would improve the battery life of these devices. Generally, approaches to improve the battery life focus on improving the speaker design and the

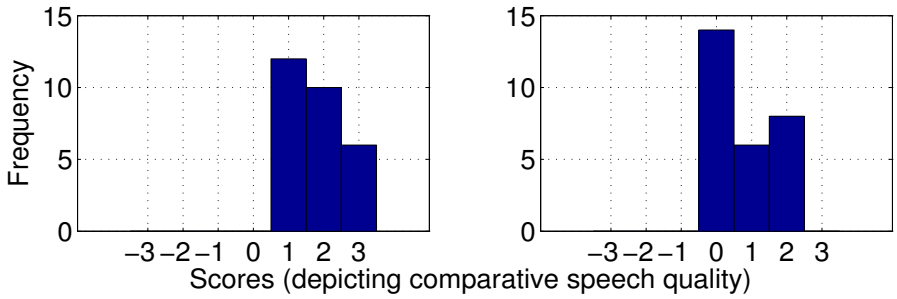

Figure 8.6: Histogram showing the frequency distribution of the CMOS scores comparing the quality of (a) narrowband telephone speech and wideband speech, and (b) narrowband telephone speech and estimated wideband speech.

efficiency of power amplifiers to reduce the current drain. The loudness enhancement is also of interest for situations that involve large gatherings such as auditoriums, where the concern is to make the speaker's voice audible to all the listeners in the gathering.

Loudness is the human perception of intensity, and is a function of the sound intensity, frequency and quality [126]. Intensity is the amount of energy flowing across a unit area surface in a second. The loudness of voices can also be enhanced by directly processing the speech signal [127]. This approach would help conserve the energy, and save power. Authors of [127] proposed a method to increase the perceived loudness without increasing the signal energy. A warped filter was used to apply a nonlinear bandwidth expansion to the formant regions of vowels on a critical band scale. As an alternative to the warped filter approach, the work proposed in this chapter exploits the excitation source characteristics of speakers which are associated with the loudness in their voice. Specifically, the voiced segments of the excitation source characteristics of soft voices are processed for enhancing the perceived loudness.

The rate of closure of the vocal folds during the quasi-periodic vibrations (associ-

ated with the articulation of voiced sounds) plays a significant role in determining the perceived loudness of a person's voice. The more rapidly the vocal folds come together, the louder the voice sounds. The steepness of the roll-off of the glottal closure instants in the LP residual signal corresponds to the rate of closure of the vocal folds. Soft voices have a gradual roll-off, while loud voices have a steep roll-off.

In order to enhance the loudness of soft voices, the LP residual signal, $r_s(n)$, of the soft voice is modified to achieve a rapid roll-off at the glottal closure instants. This is achieved by modifying the Hilbert envelope, $h_s(n)$, of $r_s(n)$ as

$$h_l(n) = h_s(n)^{\alpha(n)}, \tag{8.3}$$

where $h_l(n)$ is the corresponding Hilbert envelope of the LP residual of the loudness enhanced speech, and $\alpha(n)$ is given by

$$\alpha(n) = \begin{cases} 0.5\left(1 + cos(\frac{h_e(n)-\psi_{min}}{\psi_{max}-\psi_{min}})\pi\right), & \psi_{min} \leq h_s(n) \leq \psi_{max} \\ 1, & h_s(n) < \psi_{min} \\ 0, & h_s(n) > \psi_{max}, \end{cases} \tag{8.4}$$

where ψ_{min} and ψ_{max} are constants that determine the steepness of the roll-off, and are suitably adjusted to improve the perceived loudness while minimizing the distortion in the resulting speech. The modified LP residual, $r_l(n)$, is given by

$$r_l(n) = r_s(n)h_l(n). \tag{8.5}$$

The loudness enhanced speech is synthesized using $r_l(n)$. Fig. 8.7 shows waveforms

of two segments of the soft voice and the corresponding loudness enhanced voice. The roll-off of the glottal closure instants is steeper in the loudness enhanced voice compared to the soft voice, which is desired. The enhanced voice sounds louder compared to the soft voice, with minimal perceivable distortions.

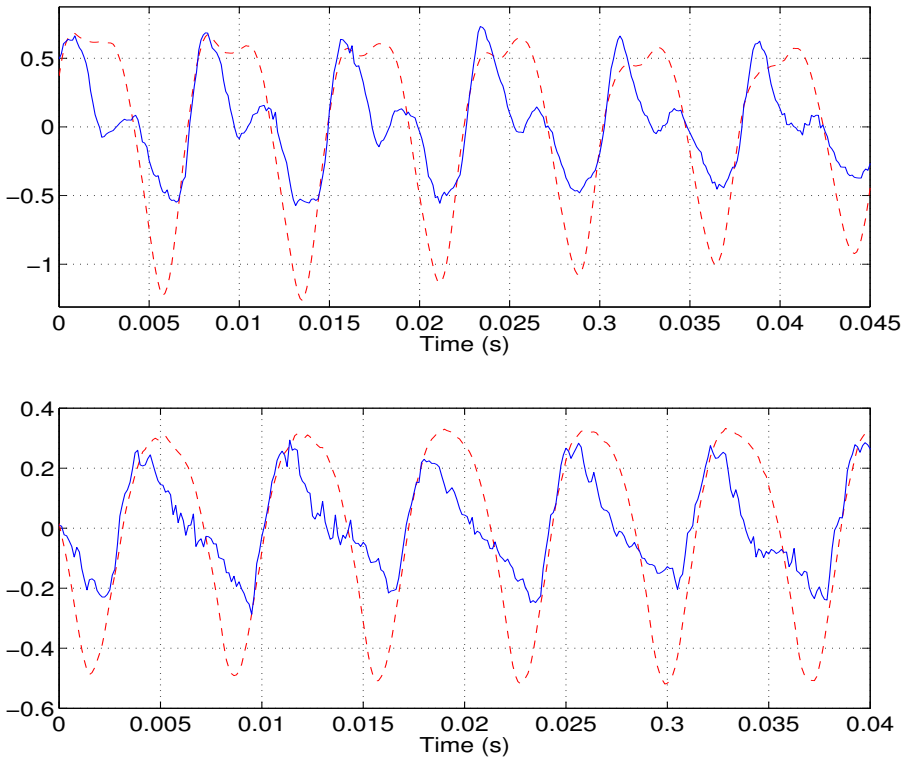

Figure 8.7: The acoustic waveforms of two different segments of soft voice (dashed line) and the loudness enhanced voice (solid line).

The improvement in the perceived loudness is evaluated using the CMOS listening test, as explained in Section 5.7. Here, the comparative rating is based on evaluating how much louder one sound is perceived compared to the other. The results of the evaluation are shown in Fig. 8.8 in the form of histograms. The listeners also considered the distortion (due to processing) in the enhanced speech during evaluation. The results show that the listeners preferred the enhanced speech over the soft voice,

with a CMOS of 1.82. Further investigation is needed to reduce the mild (but, not annoying) distortions in the enhanced speech.

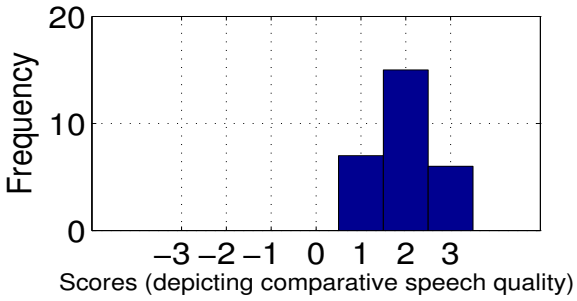

Figure 8.8: Histogram showing the frequency distribution of the CMOS score comparing the quality of the soft voice and the corresponding loudness-enhanced voice.

8.4 SUMMARY

In this chapter, the techniques used to improve the perceptual quality of the throat microphone speech was extended for two applications, namely, bandwidth extension of the narrowband telephone speech and loudness enhancement of soft voices. The bandwidth extension of the narrowband telephone speech involved the recovery of the wideband spectral envelope and the regeneration of the wideband residual signal. Spectral mapping technique was used for the recovery of the wideband spectral envelope, and the voiced segments of the narrowband residual were modified (to reduce distortion) prior to regeneration of the wideband residual signal. Subjective evaluation showed that the approach slightly improves the perceptual quality of the narrowband speech. For the loudness enhancement of soft voices, a technique that enhanced the loudness by modifying the LP residual (especially around the instants of significant excitation) was proposed. The enhanced speech was rated as perceptually louder com-

pared to the soft voice. A perceivable, but not annoying, distortion was present in the enhanced speech. Improvement in the technique is required for better results. The advantage of the technique, both for bandwidth extension and loudness enhancement, is that it does not contribute any annoying artifact to the enhanced speech.

CHAPTER 9

SUMMARY AND CONCLUSIONS

9.1 SUMMARY

In this thesis, the focus was on processing the speech signals obtained from the throat microphone (TM) for improving its perceptual quality and using it for developing automatic speech systems. The perceived lack of naturalness in the TM speech was reduced by exploiting the characteristics of the high quality normal microphone (NM) speech. The intelligibility of the TM speech in noisy ambience was exploited for speech applications in noisy conditions. Techniques used to improve the perceptual quality of the TM speech were extended for two applications: (1) bandwidth extension of the narrowband speech, and (2) loudness enhancement of soft voices.

The throat microphone is a skin vibration transducer that is placed close to the vocal folds. It is a preferred choice for use in adverse conditions as it is relatively insusceptible to the noisy ambience. It records speech that is intelligible even in noisy conditions. However, the speech from a throat microphone sounds unnatural, unlike the high quality speech recorded from a normal microphone. Improving the perceptual quality of the TM speech would help alleviate the discomfort that may arise due to prolonged listening of the TM speech.

The perceptual quality of a speech signal depends on the acoustic characteristics [79]. Hence, the differences in the perceptual quality of the speech signals from the throat microphone and the normal microphone depend on the differences in the acous-

tic characteristics of the two speech signals. The acoustic characteristics of various sound units in the TM and NM speech were studied. This study showed that acoustic differences between the TM and NM speech exist both in the vocal tract characteristics and the excitation source characteristics for different sound units.

Some of the spectral characteristics that distinguish the TM speech from the NM speech are: (1) The absence of some of the higher frequencies, clearly seen in the voiced sounds and fricatives, (2) the presence of distinct formant-like structures during the closure-phase of voiced stop consonants (unlike the low frequency voice bar which is similar for all voiced stops in the case of NM speech), and (3) the presence of oral resonances in the nasal consonants. The excitation source characteristics derived from the TM and NM speech differ in the modifications imposed by the vocal tract system on the strength of the instants of significant excitation for voiced sounds. In the NM speech, the strength of the instants is comparatively high for vowels and low for voiced consonants. In contrast, in the TM speech, the strength of the instants is comparable for the vowels and the voiced consonants.

The task of enhancing the TM speech exploited the characteristics of the high perceptual quality of the NM speech. To compensate for the acoustic differences, the task was divided into two subtasks: (1) estimating the spectral features of the NM speech, given that of the TM speech, and (2) modifying the TM residual so as to obtain the necessary modification in the strength of the instants, similar to the modifications present in the strength of the instants in the NM residual. The first subtask involved mapping of the spectral features (wLPCC) of the TM speech onto the spectral features of the NM speech. The nonlinear mapping property of the MultiLayered Feed Forward

Neural Network (MLFFNN) was used for the spectral mapping. The proposed autocorrelation method to derive LP coefficients from wLPCCs guarantees the stability of the all-pole synthesis filter, which is essential for distortion free speech. The mapping was effective for most of the sound units, with the exception of the unvoiced fricatives.

The second subtask of modifying the voiced segments of the TM residual involved the emphasis of the instants in the vowel segments, and deemphasis of the instants in the voiced consonant segments. The modification required the automatic identification of the broad phoneme category of each of the voiced sounds. This identification followed by modification was done by (a) mapping the parameters (normalized error, gross spectral features and log frame energy, which help to distinguish between the voiced segments, derived from each frame are concatenated to form the feature vector) of the TM speech onto the corresponding parameters of the NM speech, and (b) using the estimated NM features for suitable modification of the TM residual. The desired modification of the strength of instants was achieved for most of the voiced sounds. The modified residual was used to excite the synthesis filter constructed using the estimated LP coefficients to obtain the enhanced speech.

Intelligibility of the TM speech was exploited for developing an automatic syllable recognizer. Performance of the TM speech based syllable recognizer was comparable to that of the NM-based syllable recognizer for most of the sound units. In general, the performance of the TM speech based system was poorer for vowels due to confusability between the front and back vowels, but better for stop consonants due to the distinct spectral features associated with the closure phase of these consonants. The robustness of the TM speech to noise was exploited to develop speaker recognition and language

identifications systems. The performance of the TM speech based systems under noisy conditions was similar to the performance under clean conditions.

The techniques used to enhance the perceptual quality of the TM speech was extended for two other applications, namely, bandwidth extension of telephone speech and loudness enhancement of soft voices. The telephone speech is band-limited between 300 Hz and 3400 Hz. The task of improving the perceptual quality of the narrowband telephone speech involved (1) mapping the narrowband spectra to the wideband spectra using the MLFFNN, and (2) modifying the voiced segments of the narrowband residual to reduce distortion, followed by regeneration of wideband residual using spectral folding. Loudness enhancement of soft voices relied on manipulating the roll-off of the instants of significant excitation of the LP residual of soft voices.

9.2 MAJOR CONTRIBUTIONS OF THE WORK

The significant contribution reported in this thesis is the attempt to improve the perceptual quality of the throat microphone speech. The TM speech and the NM speech differ, both in the vocal tract characteristics and the excitation source characteristics. Therefore, the improvement involves compensating the differences in these characteristics.

Major contributions of this thesis are:

- A detailed acoustic analysis of the sound units in the TM speech.
- A method to effectively map the spectral features of the TM speech onto the spectral features of the NM speech.
- An approach that uses the autocorrelation method to derive LP coefficients from

wLPCCs. This approach guarantees the stability of the synthesis filter.

- A method that modifies the strength of the instants of voiced segments (emphasizing those of vowels and deemphasizing those of voiced consonants) of the TM residual to obtain an estimate of the modification of the strengths of instants seen in the NM residual. This involves an automatic discrimination among the voiced sounds using minimal speech data.
- An HMM based syllable recognizer using only the TM speech for isolated syllable recognition and group classification.
- Robust speaker recognition and language identification systems whose performance does not degrade in noisy conditions.
- An effective mapping of the narrowband telephone speech onto the wideband normal speech. The distortions in the telephone speech are reduced by performing a relative emphasis of the high SNR segments of the LP residual with respect to the low SNR segments.
- A method to enhance the loudness of soft voices which manipulates the LP residual around the instants of glottal closure, without distorting the enhance speech.

9.3 DIRECTIONS FOR FUTURE RESEARCH

- The spectral mapping network is unable to effectively map the spectra of the fricatives due to random energy (noise-like) distribution in these sounds. Hence an approach to suitably map the fricatives is necessary.

- The spectral mapping technique described in Chapter 4 is speaker-specific. In a speaker specific mapping, the formant locations and bandwidth will be the same in both the TM speech spectra and the NM speech spectra. If however, a speaker-independent mapping is used, then the estimated NM speech spectra may not represent the vocal tract characteristics of the target speaker. Differences in the formant locations may cause distortion in the synthesized speech. Hence, approaches to achieve a speaker independent mapping need to be explored.

- In the process of modifying the TM residual, the ratio of the normalized errors of the TM and NM speech is used to emphasize/deemphasize the strength of the instants in the vowels/voiced consonant segments. A robust feature that represents the strength of the instants can be explored to achieve a more effective modification. The characteristics of the nasal sounds of the NM speech are not effectively estimated, both in spectral mapping as well as residual modification. The spectral mapping is less effective for nasals than for other voiced sounds. This is because the all-pole source filter model (LP) used does not compensate for the zeros which may be introduced due to nasal coupling. In the residual modification too, some of the nasal segments are not sufficiently deemphasized. Hence, an approach to effectively estimate the characteristics of the sounds in the NM speech is necessary.

- The TM speech was used for developing a syllable-recognizer. This could be used for command-and-control applications. The study could be extended for

recognizing continuous speech, especially in noisy conditions. The robustness of the TM speech to noise was exploited for speaker recognition and language identification applications. However the TM speech based systems may degrade if the surrounding noise is highly vibratory in nature. Hence, approaches for processing the TM speech in such conditions need to be explored. The feasibility of using the TM speech for speech encoding can also be explored.

BIBLIOGRAPHY

[1] D. W. Martin, "Magnetic throat microphones of high sensitivity," *J. Acoust. Soc. Amer.*, vol. 19, pp. 43–50, Jan. 1947.

[2] J. R. Deller, J. G. Proakis and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals.* New York: Macmillan Publishing Company, 1993.

[3] K. M. Keenaghan, *A Novel Non-Acoustic Voiced Sensor: Experimental Results and Characterization.* MS thesis, Electrical and Computer Engg., Worcester Polytechnic Institute, Feb. 2004.

[4] B. Iser and G. Schmidt, "Bandwidth extension of telephony speech," *EURASIP Newsletter*, vol. 16, pp. 2–24, June 2005.

[5] P. Satyanarayana, *Short Segment Analysis of Speech for Enhancement.* PhD thesis, Dept. Comp. Sci. Engg, Indian Institute of Technology Madras, Chennai, Feb. 1999.

[6] R. J. Jones, S. Downey, and J. S. Mason, "Continuous speech recognition using syllables," in *Proc. Eur. Conf. Speech Comm. Technol.*, (Rhodes, Greece), pp. 1171–1174, Sept. 1997.

[7] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, "High performance robust speech recognition using stereo training data," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, (Salt Lake City, Utah, USA), pp. 301–304, May 2001.

[8] B. S. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 64, pp. 460–475, Apr. 1976.

[9] J. P. Campbell, Jr, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, pp. 1437–1462, Sept. 1997.

[10] Y. K. Muthusamy, E. Bardnard, and R. A. Cole, "Reviewing automatic language identification," *IEEE Signal Processing Magazine*, vol. 11, pp. 33–41, Oct. 1994.

[11] M. A. Zissman, "Overview of current techniques for automatic language identification of speech," in *IEEE Intl Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 60–62, Dec. 1995.

[12] D. R. Brown III, K. Keenaghan, and S. Desimini, "Measuring glottal activity during voiced speech using a tuned electromagnetic resonating collar sensor," *J. Measurement Science and Tech.*, vol. 16, pp. 2381–2390, 2005.

[13] T. F. Quatieri, D. Messing, K. Brady, and W. B. Campbell, "Exploiting nonacoustic sensors for speech enhancement," in *Workshop on multimodal user authentication (MMUA)*, (Santa Barbara, CA), pp. 66–73, Dec. 2003.

[14] L. C. Ng, G. C. Burnett, J. F. Holzrichter, and T. J. Gable , "Denoising of human speech using combined acoustic and EM sensor signal processing," in *Proc. IEEE*

Int. Conf. Acoust., Speech and Signal Processing (ICASSP), vol. 1, (Istanbul Turkey), pp. 229–232, June 2000.

[15] Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, and Xuedong Huang, "Air- and bone-conductive integrated microphones for robust speech detection and enhancement," in *IEEE Intl. Workshop on Automatic Speech Recognition and Understanding (ASRU)*, (Virgin Islands, USA), pp. 249–254, Dec. 2003.

[16] J. Hershey, T. Kristjansson, Z. Zhang, "Model-based fusion of bone and air sensors for speech enhancement and robust speech recognition," in *Proc. ISCA Workshop on Statistical and Perceptual Audio Processing*, (Jeju Island, Korea), Oct. 2004.

[17] T. F. Quatieri, K. Brady, D. Messing, J. P. Campbell, W. M. Campbell, M. S. Brandstein, C. J. Weinstein, J. D. Tardelli and P. D. Gatewood , "Exploiting nonacoustic sensors for speech encoding," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 14, pp. 533–544, Mar. 2006.

[18] P. Heracleous, Y. Nakajima, H. Saruwatari, and Kiyohiro Shikano, "A tissue-conductive acoustic sensor applied in speech recognition for privacy," in *Proc. Jt.Conf. Smart Objects and Ambient Intelligence*, (Grenoble, France), pp. 93–97, Oct. 2005.

[19] Z. Zhang, Z. Liu, M. Sinclair, A. Acero, L. Deng, J. Droppo, X. Huang, and Y. Zeng, "Multi-sensory microphones for robust speech detection, enhancement and recognition," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, (Montreal, Quebec, Canada), pp. 781–784, May 2004.

[20] J. Droppo, A. Acero, and L. Deng, "Evaluation of splice on aurora 2 and 3 tasks," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, (Denver, Colorado), pp. 29–32, Sept. 2002.

[21] Z. Liu, Z. Zhang, A. Acero, J. Droppo, and X. Huang, "Direct filtering for air- and bone-conductive microphones," in *Proc. Intl. Workshop on Multimedia Signal Processing (MMSP)*, (Siena, Italy), Sept. 2004.

[22] Z. Liu, A. Subramanya, Z. Zhang, J. Droppo, and A. Acero, "Leakage model and teeth clack removal for air- and bone-conductive integrated microphones," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, (Philadelphia, PA, USA), pp. 1093–1096, Mar. 2005.

[23] A. Subramanya, L. Deng, Z. Liu, and Z. Zhang, "Multi-sensory speech processing: Incorporating automatically extracted hiden dynamic information," in *IEEE Int. Conf. Multimedia and Expo (ICME)*, (Amsterdam), July 2005.

[24] A. Subramanya, Z. Zhang, Z. Liu, J. Droppo, and A. Acero, "A graphical model for multisensory speech processing in air-and-bone conductive integrated microphones," in *Proc. Eur. Conf. Speech Comm. Technol.*, (Lisbon, Portugal), pp. 393–396, Sept. 2005.

[25] R. Hu and D. V. Anderson, "Single acoustic-channel speech enhancement based on glottal correlation using non-acoustic sensor," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, (Jeju Island, Korea), pp. 861–864, Oct. 2004.

[26] C. Demiroglu, *Multisensor Segmentation-based Noise Suppression for Intelligibility Improvement in MELP Coders.* PhD thesis, School of Electrical and Comp. Engg, Georgia Institute of Technology, May. 2006.

[27] C. Demiroglu and D. V. Anderson, "A soft decision mmse amplitude estimator as a noise preprocessor to speech coders using a glottal sensor," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, (Jeju Island, Korea), pp. 857–860, Oct. 2004.

[28] R. Hu and D. V. Anderson, "Improved perceptually inspired speech enhancement using a psychoacoustical model," in *Thirty-Eighth Asilomar Conf. Signals, Systems and Computers*, vol. 1, (Pacific Grove, California), pp. 440–444, Nov. 2004.

[29] J. F. Holzrichter, "Methods and apparatus for non-acoustic speech characterization and recognition," (http://www.freepatentsonline.com/6006175.html), 1999.

[30] K. Saenko, T. Darrell, and J. Glass, "Articulatory features for robust visual speech recognition," in *Proc. ICMI*, (Pennsylvania, USA), pp. 152–158, Oct. 2004.

[31] S. Roucas and V. Viswanathan, "Word recognition using multisensor speech input in high ambient noise," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, pp. 868–871, Apr. 1986.

[32] M. Graciarena, H. Franco, K. Sonmez, and H. Bratt, "Combining standard and throat microphones for robust speech recognition," *IEEE Signal Processing Letters*, vol. 10, pp. 72–74, Mar. 2003.

[33] M. Graciarena, H. Franco, G. Myers, C. Cowan, F. Cesari, and V. Abrash, "Combination of standard and throat microphones for robust speech recognition in hihgly noisy environments," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, (Jeju Island, Korea), pp. 809–812, Oct. 2004.

[34] S.-C. Jou, T. Schultz, and A. Waibel, "Adaptation for soft whisper recognition using a throat microphone," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, (Jeju Island, Korea), pp. 1493–1496, Oct. 2004.

[35] C. Demiroglu and D. V. Anderson, "Broad phoneme class recognition in noisy environments using the GEMS device," in *Thirty-Eighth Asilomar Conf. Signals, Systems and Computers*, vol. 1, (Pacific Grove, California), pp. 1805–1808, Nov. 2004.

[36] C. Demiroglu and D. V. Anderson, "Noise robust digit recognition using a glottal radar sensor for voicing decision," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, (Jeju Island, Korea), pp. 813–816, Oct. 2004.

[37] P. S. Aleksic and A. K. Katsaggelos, "An audio-visual person identification and verification system using faps as visual features," in *Workshop on multimodal user authentication*, (Santa Barbara, CA), pp. 80–84, Dec. 2003.

[38] W. B. Campbell, T. F. Quatieri, J. P. Campbell, and C. J. Weinstein, "Multimodal speaker authentication using nonacoustic sensors," in *Workshop on multimodal user authentication (MMUA)*, (Santa Barbara, CA), pp. 215–222, Dec. 2003.

[39] L. C. Ng, T. J. Gable, and J. F. Holzrichter, "Speaker verification using combined acoustic and EM sensor signal processing," (http://speech.llnl.gov/), 2001.

[40] J. F. Holzrichter and L. C. Ng, "Speech coding using em sensor and acoustic signals," in *IEEE Proc. Digital Signal Processing Workshop, and the 2nd Signal Processing Education Workshop*, vol. 10, (Pine Mountain, GA), 2002.

[41] S. Dusan, J. Flanagan, A. Karve, and M. Balaraman, "Speech coding using trajectory compression and multiple sensors," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, (Jeju Island, Korea), pp. 1993–1996, Oct. 2004.

[42] H. Hermansky, C. Avendano, and E. A. Wan, "Noise reduction and recovery of missing frequencies in speech," in *Proc. of 15th Annual Speech Research Symposium*, (Baltimore MD), pp. 124–129, 1995.

[43] G. Miet, A. Gerrits, and J. C. Valire, "Low-band extension of telephone-band speech," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, vol. 3, (Istanbul), pp. 1851–1854, June 2000.

[44] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter, "Speech enhancement via frequency bandwidth extension using line spectral frequencies," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, vol. 1, (Salt Lake City, Utah, USA), pp. 665 – 668, May 2001.

[45] S. Jaisimha and Y. Soon, "Bandwidth extension of narrow band speech using cepstral linear prediction," in *Proc. Int. Conf. on Info., Comm. and Signal Processing-Pacific Rim Conf. on Multimedia*, (Singapore), pp. 1404–1407, Dec. 2003.

[46] C. Avendano, H. Hermansky, and E. A. Wan, "Beyond Nyquist: Towards the recovery of broad-bandwidth speech from narrow-bandwidth speech," in *Proc. Eur. Conf. Speech Comm. Technol.*, (Madrid, Spain), pp. 165–168, Sept. 1995.

[47] J. A. Fuemmeler, R. C. Hardie, and W. R. Gardner, "Techniques for the regeneration of wideband speech from narrowband speech," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 266–274, 2001.

[48] N. Enbom and W. B. Kleijn, "Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients," in *Proc. of the IEEE Workshop on Speech Coding*, (Porvoo, Finland), pp. 171–173, Sept. 1999.

[49] J. Epps and W. H. Holmes, "A new technique for wideband enhancement of coded narrowband speech," in *Proc. of the IEEE Workshop on Speech Coding*, (Porvoo, Finland), pp. 174–176, Sept. 1999.

[50] R. Hu, V. Krishnan, and D. V. Anderson, "Speech bandwidth extension by improved codebook mapping towards increased phonetic classification," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, (Lisbon, Portugal), pp. 1501–1504, Sept. 2005.

[51] H. Ehara, T. Morii, M. Oshikiri, K. Yoshida, and K. Honma, "Design of bandwidth scalable LSF quantization using interframe and intraframe prediction," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, (Lisbon, Portugal), pp. 1493–1496, Sept. 2005.

[52] Y. M. Cheng, D. O'Shaughnessy, and P. Mermelstein, "Statistical recovery of wideband speech from narrowband speech," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 544–548, Oct. 1994.

[53] K.-Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM-based transformation," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, vol. 3, (Istanbul, Turkey), pp. 1847–1850, June 2000.

[54] M. L. Seltzer, A. Acero, and J. Droppo, "Robust bandwidth extension of noise-corrupted narrowband speech," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, (Lisbon, Portugal), pp. 1509–1512, Sept. 2005.

[55] G. Chen and V. Parsa, "HMM-based frequency bandwidth extension for speech enhancement using line spectral frequencies," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, vol. 1, (Montreal, Quebec, Canada), pp. 709–712, May 2004.

[56] P. Jax and P. Vary, "Wideband extension of telephone speech using a hidden markov model," in *Proc. of the IEEE Workshop on Speech Coding*, (Delavan, WI, USA), pp. 133–135, Sept. 2000.

[57] Y. Qian and P. Kabal, "Combining equalization and estimation for bandwidth extension of naarrowband speech," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, (Montreal, Quebec, Canada), pp. 713–716, May 2004.

[58] B. Iser and G. Schmidt, "Neural networks versus codebooks in an application for bandwidth extension of speech signals," in *Proc. Eur. Conf. Speech Comm. Technol.*, (Geneva, Switzerland), pp. 565–568, Sept. 2003.

[59] A. Uncini, F. Gobbi, and F. Piazza, "Frequency recovery of narrow-band speech using adaptive spline neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, vol. 2, (Phoenix, Arizona, USA), pp. 997–1000, Mar. 1999.

[60] W.-S. Hsu, *Robust Bandwidth Extension of Narrowband Speech.* M.E thesis, Dept. Elect. and Comp. Sci. Engg, McGill University, Montreal, Canada, Nov 2004.

[61] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, vol. 4, (Washington DC), pp. 428–431, Apr. 1979.

[62] J. P. Cabral and L. C. Oliveira, "Pitch-synchronous time-scaling for high-frequency excitation regeneration," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, (Lisbon, Portugal), pp. 1513–1516, Sept. 2005.

[63] J. Harrington, *Aspects of Speech Technology (Chapter 2).* Edinburg: Edinburg University Press, 1988.

[64] K. N. Stevens, "Acoustic correlates of some phonetic categories," *J. Acoust. Soc. Amer.*, vol. 63, pp. 836–842, Sept. 1980.

[65] A. S. Abramson, P. W. Nye, J. B. Henderson, and C. W. Marshall, "Vowel height and the perception of consonant nasality," *J. Acoust. Soc. Amer.*, vol. 70, pp. 329–339, Aug. 1980.

[66] A. S. House, "On vowel duration in English," *J. Acoust. Soc. Amer.*, vol. 33, pp. 1174–1178, Sept. 1961.

[67] S. Hawkins and K. N. Stevens, "Acoustic and perceptual correlates of the non-nasal-nasal distiction for vowels," *J. Acoust. Soc. Amer.*, vol. 77, pp. 1560–1575, Apr. 1985.

[68] M. Halle, G. W. Hughes, and J. P. A. Radley, "Acoustic properties of stop consonants," *J. Acoust. Soc. Amer.*, vol. 29, pp. 107–116, Jan. 1957.

[69] K. N. Stevens and S. E. Blumstein, "Invariant cues for place of articulation in stop consonants," *J. Acoust. Soc. Amer.*, vol. 64, pp. 1358–1368, Nov. 1978.

[70] D. Kewley-Port, "Measurement of formant transitions in naturally produced stop consonant-vowel syllables," *J. Acoust. Soc. Amer.*, vol. 72, pp. 379–389, Aug. 1982.

[71] O. Fujimura, "Analysis of nasal consonants," *J. Acoust. Soc. Amer.*, vol. 34, pp. 1865–1875, Dec. 1962.

[72] J. M. Heinz and K. N. Stevens, "On the properties of voiceless fricative consonants," *J. Acoust. Soc. Amer.*, vol. 33, pp. 589–596, May 1961.

[73] G. W. Hughes and M. Halle, "Spectral properties of fricative consonants," *J. Acoust. Soc. Amer.*, vol. 28, pp. 303–310, Mar. 1956.

[74] B. E. F. Lindblom and J. E. F. Sundberg, "Acoustical consequences of lip, tongue, jaw and larynx movement," *J. Acoust. Soc. Amer.*, vol. 50, pp. 1166–1178, May 1971.

[75] M. V. Mathews, J. E. Miller, and E. E. David, "Pitch synchronus analysis of voiced sounds," *J. Acoust. Soc. Amer.*, vol. 33, pp. 179–186, Feb. 1961.

[76] B. Yegnanarayana, "Formant extraction from linear-prediction phase spectra," *J. Acoust. Soc. Amer.*, vol. 63, pp. 1638–1640, May 1978.

[77] B. Yegnanarayana, "On timing in time-frequency analysis of speech signals," *Sadhana*, vol. 21, pp. 5–20, Feb. 1996.

[78] P. Lieberman and S. E. Blumstein, *Speech physiology, speech perception, and acoustic phonetics.* Cambridge, Great Britain: Cambridge University Press, 1988.

[79] Peter Ladefoged, *A Course in Phonetics.* Orlando, FL, U. S. A: Harcourt College Publishers, 2001.

[80] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ: Prentice Hall, 1993.

[81] J .D. Markel and A. H. Gray, Jr, *Linear Prediction of Speech.* Berlin: Springer Verlag, 1976.

[82] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[83] M. R. Schroeder, "Linear prediction, extremal entropy and prior information in speech signal analysis and synthesis," *Speech Comm.*, vol. 1, no. 1, pp. 9–20, 1982.

[84] B. Yegnanarayana, *Artificial Neural Networks.* New Delhi: Prentice Hall India, 1999.

[85] Hemant Misra, M. Shajith Ikbal, and B. Yegnanarayana, "Speaker-specific mapping for text-independent speaker recognition," *Speech Comm.*, vol. 39, pp. 301–310, Feb. 2003.

[86] S. Haykin, *Neural Networks: A Comprehensive Foundation.* NJ: Prentice-Hall International, 1999.

[87] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, pp. 251–257, 1991.

[88] K. Funahashi, "On the approximate realization of continuous mapping by neural networks," *Neural Networks*, vol. 2, no. 3, pp. 183–192, 1989.

[89] J. Makhoul, A. El-Jaroudi, and R. Schwartz, "Partitioning capabilities of two layer neural networks," *IEEE Trans. Signal Processing*, vol. 39, pp. 1435–1440, June 1991.

[90] E. D. Sontag, "Feedback stabilization using two-hidden-layer nets," *IEEE Trans. Neural Networks*, vol. 3, pp. 981–990, Nov. 1992.

[91] E. D. Sontag, "Feedforward nets for interpolation and classification," *J. Computing System Sciences*, vol. 45, pp. 20–48, 1992.

[92] M. S. Ikbal, H. Misra, and B. Yegnanarayana, "Analysis of autoassociative mapping neural networks," in *Proc. IEEE Int. Joint Conf. Neural Networks*, (Washington), pp. 854–858, July 1999.

[93] R. P. Lippmann, "An introduction to computing with neural nets," *IEEE ASSP Magazine*, vol. 2, pp. 4–22, Apr. 1987.

[94] S. R. M. Prasanna, *Event-Based Analysis of Speech.* PhD thesis, Dept. Comp. Sci. Engg, Indian Institute of Technology, Madras, Mar. 2004.

[95] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, pp. 309–319, Aug. 1979.

[96] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing.* New Jersey: Prentice Hall, 2000.

[97] P. Eswar, S. K. Gupta, C. Chandrasekhar, B. Yegnanarayana, and K. N. Reddy, "An acoustic-phonetic expert for analysis and processing of continuous speech in Hindi," in *Proc. European Conf. Speech Technology*, (Edinburgh, UK), pp. 369–372, Sept. 1987.

[98] S. Greenberg, "Speaking in shorthand - A Syllable-centric perspective for understanding pronunciation variation," *Speech Comm.*, vol. 29, pp. 159–176, Nov. 1999.

[99] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski, and G. R. Dodington, "Syllable-based large vocabulary continuous speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 358–366, May 2001.

[100] A. Tsopanoglou and N. Fakotakis, "Selection of the most effective set of subword units for a HMM-based speech recognition system," in *Proc. Eur. Conf. Speech Comm. Technol.*, vol. 3, (Rhodes, Greece), pp. 1231–1234, Sept. 1997.

[101] C. Chandrasekhar, *Neural network models for recognition of Stop Consonant-Vowel (SCV) segments in continuous speech.* PhD thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Apr. 1996.

[102] S. V. Gangashetty, C. Chandra Sekhar, and B. Yegnanarayana, "Extraction of fixed dimension patterns from varying duration segments of consonant-vowel utterances," in *Proc. IEEE Int. Conf. Intelligent Sensing and Information Processing*, (Chennai, India), pp. 159–164, Jan. 2004.

[103] S. V. Gangashetty, K. S. Rao, A. N. Khan, C. Chandrasekhar, and B. Yegnanarayana, "Combining evidence from multiple modular networks for recognition of consonant-vowel units of speech," in *Proc. IEEE Int. Joint Conf. Neural Networks*, vol. 1, (Portland, Oregon, USA), pp. 686–691, July 2003.

[104] S. V. Gangashetty, C. Chandrasekhar, and B. Yegnanarayana, "Constraint satisfaction model for enhancement of evidence in recognition of consonant-vowel utterances," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, vol. 3, (Maryland, USA), pp. 201–204, July 2003.

[105] A. N. Khan, "Recognition of syllable like units in Indian languages," in *Proc. Int. Conf. Natural Language Processing*, (Mysore India), pp. 135–141, Dec. 2003.

[106] S. V. Gangashetty, *Neural network models for Recognition of Consonant-Vowel Units of Speech in Multiple Languages.* PhD thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Feb. 2005.

[107] Y.-Q. Bao, L. Zhao, C.-R. Zou, "Study on speaker recognition under noise environments based on PCANN," in *Proc. Int. Conf. Machine Learning and Cybernetics*, (Shangai, China), pp. 3770–3774, Aug. 2004.

[108] F. J. Goodman, A. F. Martin, and R. E. Wohlford, "Improved automatic language identification in noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, pp. 528–531, May 1989.

[109] D. B. Fry, *The Physics of Speech.* Cambridge, Great Britain: Cambridge University Press, 1979.

[110] D. O'Shaughnessy, "Speaker recognition," *IEEE ASSP Magazine*, vol. 3, pp. 4–17, Oct. 1986.

[111] G. R. Doddington, "Speaker recognition-Identifying people by their voices," *Proc. IEEE*, vol. 73, pp. 1651–1664, Nov. 1985.

[112] B. Yegnanarayana, K. S. Reddy, and S. P. Kishore, "Source and system features for speaker recognition using AANN models," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, (Salt Lake City, Utah, USA), pp. 409–412, May 2001.

[113] M. Faundez-Zanuy and D. Rodriquez-Parcheron, "Speaker recognition using residual signal of linear and nonlinear prediction models," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, vol. 2, (Sydney, Australia), pp. 121–124, Dec. 2004.

[114] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 569–586, Sept. 1999.

[115] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, pp. 52–55, Jan. 2006.

[116] Ch. S. Gupta, *Significance of Source Features for Speaker Recognition.* MS thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Apr. 2003.

[117] M. S. Ikbal, *Autoassociative Neural Network Model for Speaker Verification.* MS thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, May. 1999.

[118] B. Yegnanarayana and S. P. Kishore, "AANN: An alternative to GMM for pattern recognition," *Neural Networks*, vol. 15, pp. 459–469, Apr. 2002.

[119] K.-P. Li, "Automatic language identification using syllabic spectral features," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, pp. 297–300, Apr. 1994.

[120] T. Nagarajan, *Implicit Systems for Spoken Language Identification.* PhD thesis, Dept. Comp. Sci. Engg, Indian Institute of Technology, Madras, Jan. 2004.

[121] L. Mary, *Multilevel Implicit Features for Language and Speaker Recognition.* PhD thesis, Dept. Comp. Sci. Engg, Indian Institute of Technology, Madras, Jan. 2007.

[122] Y. K. Muthusamy and R. A. Cole, "Automatic segmentation and identification of ten languages using telephone speech," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, (Alberta, Canada), pp. 1007–1010, Oct. 1992.

[123] J. T. Foil, "Language identification using noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, pp. 861–864, Apr. 1986.

[124] L. Mary, K. S. R. Murthy, S. R. M. Prasanna, and B. Yegnanarayana, "Features for speaker and language identification," in *Proc. ODYSSEY - The Speaker and Language Recognition Workshop*, (Toledo, Spain), June 2004.

[125] M. Chaitanya, *Single Channel Speech Enhancement.* MS thesis, Dept. Comp. Sci. Engg, Indian Institute of Technology Madras, Chennai, Apr. 2005.

[126] M. A. Boillot and J. G. Harris, "A warped bandwidth expansion filter," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, (Philadelphia, PA, USA), pp. 65–68, Mar. 2005.

[127] M. A. Boillot and J .G. Harris, "A loudness enhancement technique for speech," in *Proceedings of the 2004 International Symposium on Circuits and Systems*, (Vancouver, Canada), pp. 616–618, May 2004.

LIST OF PUBLICATIONS BASED ON THE THESIS

REFEREED JOURNAL

A. Shahina and B. Yegnanarayana, "Mapping Speech Spectra from Throat Microphone to Close-Speaking Microphone: A Neural Network Approach," accepted for pulication in *EURASIP J. Adv. Signal Processing,* 2007.

REFEREED INTERNATIONAL CONFERENCES

1. A. Shahina and B. Yegnanarayana, "Mapping Neural Networks for Bandwidth Extension of Narrowband Speech," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, (Pittsburgh, Pennsylvania), Sep. 2006.
2. A. Shahina and B. Yegnanarayana, "Recognition of Consonant-Vowel Units in Throat Microphone Speech," in *Proc. Int. Conf. Natural Language Processing (ICON)*, (Indian Institute of Technology, Kanpur, India), pp. 85–92, Dec. 2005.
3. A. Shahina and B. Yegnanarayana, "Language Identification in Noisy Environments using Throat microphone Signals," in *Proc. Int. Conf. Intelligent Sensing and Information Processing (ICISIP)*, (Chennai, India), pp. 400–403, Jan. 2005.
4. A. Shahina, B. Yegnanarayana, and M. R. Kesheorey, "Throat Microphone Signal for Speaker Recognition," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, (Jeju Island, Korea), pp. 2341–2344, Oct. 2004.