IMPROVING CHILDREN'S SPEECH RECOGNITION UNDER MISMATCHED CONDITION USING ARTIFICIAL BANDWIDTH EXTENSION

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

SUNIL Y

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, INDIA

JUNE 2016

IMPROVING CHILDREN'S SPEECH RECOGNITION UNDER MISMATCHED CONDITION USING ARTIFICIAL BANDWIDTH EXTENSION

SUNIL Y

Certificate

This is to certify that the thesis entitled "**IMPROVING CHILDREN'S SPEECH RECOGNITION UNDER MISMATCHED CONDITION USING ARTIFICIAL BANDWIDTH EXTENSION**", submitted by **Sunil Y** (08610211), a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under our supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in our opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Prof. Rohit Sinha

Prof. S. R. Mahadeva Prasanna

Dept. of Electronics and Electrical Engineering

Indian Institute of Technology Guwahati

Guwahati - 781 039, India.

Date:

Acknowledgements

I am obliged to GOD for his divine guidance and blessings.

This thesis would not have been possible without the immense help and support of several people in various measures. I would like to convey my acknowledgment to all of them.

First and foremost, I express my sincere gratitude to my research supervisors, Prof. Rohit Sinha and Prof. S. R. M. Prasanna for providing me an opportunity to work under there guidance. It would be completely impossible for me to bring the research as well as the thesis to this form without the immense facilities provided by them in the EMST Laboratory and the freedom of work they have given to me.

I am thankful to my doctoral committee members Prof. S. Dandapat, Prof. P. K. Bora, Prof. C. Mahanta and Dr. Samith Bhattacharya for their encouragement and valuable suggestions on my work. I would like to thank faculty members and the office staffs of the Department of Electronics and Electrical Engineering, IIT Guwahati, for their help in carrying out this research work.

I am thankful to K.K. Ramesh, Deepak K. T, Nagaraj Adiga, Malaya Kumar Nath, Biswajit Dev Sarma, Gayadhar Pradhan, Syed Shahnawazuddin, Rohan Kumar Das, Anurag Singh, Sibasankar Padhy, Ramesh K Bhukya and all other members in the EMST Laboratory.

I would like to thank my senior members Dr. S.R. Nirmala, Dr. P. Krishnamoorthy, Dr. M. Sabarimalai Manikandan, Dr. H.S. Jayanna, Dr. D. Pati, Dr. Govind D., Dr. Gayadhar Pradhan and Dr. Haries B C. My special thanks to Dr. L. N. Sharma for maintaining the EMST laboratory smoothly.

I am thankful to my wife for her sacrifice and support. I am heartily thankful to my daughter who has sacrificed her valuable time for me.

I attribute this achievement to my parents and parents-in-law for their constant blessings, support, silent prayers for my success and making me stand in this position.

Sunil Y

Abstract

Children's speech production system distinguishes itself from the adults' by shorter vocal tract length and higher pitch value. Due to shorter vocal tract length, formant frequency values shift to higher band (3400-8000 Hz) region. The higher pitch value results in relatively more fluctuations in the spectrum as compared to adults. Narrowband (NB, 300-3400 Hz) automatic speech recognition (ASR) performance therefore degrades due the loss of higher band spectral content and fluctuating spectrum which is more significant in children's speech compared to adults' speech. The effort of this work is to develop methods that restores higher band spectral information using the mutual information between narrowband and higher band spectra. In this work, these are collectively termed as artificial bandwidth extension (ABWE) methods.

The ASR is a connected digit recognition task which has models trained using speech of adults, and tested using speech of either adults or children. The testing case with adults' speech is termed as *matched condition*, where as, testing using children's speech is termed as *mismatched condition*. The connected digit recognition task is carried out under mismatched condition for narrowband, wideband (50 - 8000 Hz), and artificially extended wideband cases. The ABWE using an existing approach significantly improves the performance of children's speech recognition under mismatched condition. It is also observed that ABWE further improves performance on top of VTLN and truncation of coefficients to minimize pitch mismatch effect, demonstrating the significance of ABWE for improving children's ASR performance under mismatched condition.

One direction for developing ABWE methods for children's speech recognition under mismatched condition is to observe the causes for the mismatch. The significant variability in the distribution of features across different classes. As a result, modelling the entire distribution using a single Gaussian mixture model (GMM) is a poor representation. Therefore class-specific modelling can be exploited for ABWE. The broad group of children (06-15 years) has significant variability in itself due to vocal tract length variation, changes in pitch values and also speaking rate. Therefore age-specific information can be exploited for ABWE. Also, the significant variability in the speaking rate can be captured in terms features that capture dynamic variability like delta features. The effectiveness of identified causes for mismatch like class specific, age specific and delta features can be first verified using statistical measures like mutual information, entropy, the ratio of mutual information to entropy and separability. The measured statistical measures indeed show significant variability between NB and higher band using the identified causes of mismatches.

The ABWE methods using class specific, age specific and delta features are developed and used in the children's speech recognition under mismatched condition. All of them show improvement in performance. A computationally efficient architecture for mel frequency cepstral coefficients (MFCC) based ABWE for ASR is developed that avoids vocoder framework for bandwidth extension. In the proposed method, the narrowband MFCC is directly converted into wideband MFCC thus avoiding the synthesis process.

Sparse representation based ABWE (SR-ABWE) algorithm is proposed using coupled dictionaries. To further enhance SR-ABWE, least square transformation has been developed to estimate wideband codes from NB interpolated codes. This is supported by the benchmark performance of look-up table mapping to estimate WB codes. Existing semi-coupled dictionary learning (SCDL) method has been explored for ABWE (SC-ABWE). This is proposed earlier for image-style-transformation application. SCDL algorithm learns bidirectional transformation iteratively with dic-

tionary learning. These dictionaries have not fully coupled and hence provide more freedom to the transformation. An improvement in the performance of SC-ABWE is observed in terms of objective quality measures. The significance of SR-ABWE is also demonstrated in children's ASR.

Keywords: children's speech, adults' speech, narrowband, wideband, ABWE, mismatched ASR, MFCC based ABWE, Sparse representation based ABWE.

Contents

List of Figures

List of Tables

List of Acronyms

ABWE	Artificial Bandwidth Extension
ABWE-AG	ABWE with Age Specific Information
ABWE-AG+Δ	ABWE with Age Specific Information and Delta features
ABWE-AG-ML	ABWE with Age Specific with Maximum Likelihood
ABWE-G	ABWE with Global Transform
ABWE-G+Δ	ABWE with Global Transform and Delta features
ABWE-GT	ABWE with Global Transform
ABWE-GT+Δ	ABWE with Global Transform and Delta features
ABWE-SUP-CLS	ABWE with Supervised Class Information
ABWE-UNSUP-CLS	ABWE with Unsupervised Class Information
AD	Adults' test set
ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
Avg.	Average
BWE	Bandwidth Extension
CC	Cepstral Coefficients
CELP	Code-excited Linear Prediction
CH	Children's test set
CMLLR	Constrained Maximum Likelihood Linear Transformation
Coupled-ABWE	Coupled Artificial Bandwidth Extension
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform

E_{HB}	Energy of Higherband
E_{LB}	Energy of Lowerband
EM	Expectation–Maximization
f_0	Fundamental Frequency
FFT	Fast Fourier Transform
GMM	Gaussian Mixture Model
HB	Higherband
HMM	Hidden Markov Model
HTK	Hidden Markov Model Toolkit
I(X;Y)	Mutual Information between feature vectors X and Y
IDFT	Inverse Discrete Fourier Transform
KSVD	K Singular Value Decomposition
LARS	Least Angle Regression
LB	Lowerband
LLR	Log-Likelihood Ratio
LP	Linear Prediction
LPC	Linear Predictive Coding
LPCC	Linear Prediction Cepstral coefficient
LR	Likelihood Ratio
LS	Least Square
LSD	Log-Spectral Distance
LSF	Line Spectral Frequencies
LST	Least Squares Transformation
MFCC	Mel-Frequency Cepstral Coefficients
MI	Mutual Information
MLLR	Maximum Likelihood Linear Regression
MLLR-COV	Maximum Likelihood Linear Regression-Covariance
MLLR-MEAN	Maximum Likelihood Linear Regression-Mean

MMSE	Minimum Mean Square Error
MOS	Mean Opinion Score
NB	Narrowband
NB-ASR	Narrowband Automatic Speech Recognition
NBI	Narrowband Interpolated
OMP	Orthogonal Matching Pursuit
PCM	Pulse-Code Modulation
PDF	Probability Density Function
PESQ	Perceptual Evaluation of Speech Quality
PSR	Primary School Reading Speech Corpus
QCQP	Quadratically Constrained Quadratic Program
RAPT	Robust Algorithm for Pitch Tracking
SC	Semi-Coupled
SC-ABWE	Semi-Coupled Artificial Bandwidth Extension
SCD	Semi-Coupled Dictionary
SCDL	Semi-Coupled Dictionary Learning
segSNR	Segmental Signal-to-Noise Ratio
SLP	Selective Linear Prediction
SLPCC	Selective Linear Prediction Cepstral Coefficients
SR	Sparse Representation
SR-ABWE	Sparse Representation Based Artificial Bandwidth Extension
STFT	Short-time Fourier Transform
SUP-CLS	Supervised Class
UNSUP-CLS	Unsupervised Class
UV	Unvoised
V	Voiced
V/UV	Voiced/Unvoised
VQ	Vector Quantization

VTL	Vocal Tract Length
VTLN	Vocal Tract Length Normalization
WB	Wideband
WB-ASR	Wideband Automatic Speech Recognition
WER	Word Error Rate
WLR	Weighted Likelihood Ratio
WSJCAM0	Wall Street Journal Cambridge University Speech Corpus
WSM	Weighted Slope Metric
WSS	Weighted Spectral Slope

1

Introduction

Contents

Most of the available automatic speech recognition (ASR) systems are developed using speech data collected from adult population. If this ASR system is tested from speech of adult speakers, then it is termed as *matched* condition. Alternatively, if the ASR system is tested using children's speech, then it is termed as *mismatched* condition. Since the structure and dynamics of the speech production systems are quite different in case of children and adult, a significant degradation in ASR performance is expected for children's speech recognition under mismatched condition. The objective of this thesis work is to develop methods for improving children's speech recognition under mismatched condition. Children's speech production system has distinguished itself from that of adults' in terms of shorter vocal tract length, higher pitch value, and slower speaking rate. Shorter vocal tract length results in higher formant frequencies, and hence the energy in the higher band (HB) is expected to be more. Because of this, children's speech requires higher bandwidth for better perception, and also improving ASR performance. The goal of the work is, given narrowband (NB) speech (0-3.4 kHz), develop methods for constructing the spectral information in the higher band (3.4-8 kHz). The process is termed as artificial bandwidth expansion (ABWE). The ABWE for children's speech is more challenging because of high variance of acoustic-phonetic parameters. The efficacy of the developed ABWE methods is demonstrated in the context of children's speech recognition under mismatched condition.

1.1 Nature of children's and adults' speech signals

The nature of children's and adults' speech signals can be observed by plotting them in different ways. Figure 1.1 plots the speech waveforms, wideband and narrowband spectrograms, short term magnitude spectra, LP spectra and smoothed spectra based on cepstral analysis for children's and adults' speech signals for the digit sequence "one three five seven". All the subplots in the first column are for the children's case and that of second second column for adults' case. Important distinctions among the two categories of speech are obvious in the different plots. The longer duration of children's speech signal indicates that the speaking rate for children is low compared to that of adults (Figures 1.1 (a) and 1.1 (b)). On an average,

the duration of each digit in children's speech is about 1.32 times of adults' speech. The distribution of spectral energy in the higher frequencies for the case of children's speech can be observed in case of wideband spectrograms (Figures 1.1 (c) and 1.1 (d)). About 40% of total spectral energy is in the higher frequency range (4-8 kHz) in case of children's speech, where as only about 10% of total energy is in higher frequency range for adults' case. The narrowband spectrograms given in Figures 1.1 (e) and 1.1 (f) illustrate about the high pitch values present in case of children's speech. The higher pitch values and also larger spectral dynamics in case of children can be observed with the help of short term spectra given in Figures 1.1 (g) and 1.1 (h). The children's speech has a pitch value of about 300 Hz where as it is about 100 Hz for adults' case. The spectral dynamics representing peak to valley in case of children's speech is about two times that of adults' case. The LP spectra plotted in Figures 1.1 (i) and 1.1 (j) shows shift in the formant values towards higher frequency range for the children's speech and hence the need for high band (HB) information for children's speech recognition. Children's speech has only 3 formants in 0-4 kHz range and adults' case has most of the formants within 0-4 kHz range. In Figures 1.1 (k) and 1.1 (l) shows cepstrally smoothed spectral plots. Solid line shows spectrum from all 26 coefs, where as dotted line shows spectrum from retaining only 13 coefficients. The fluctuations are more in case of children's speech compared to adults' speech. Significant gain in terms reducing fluctuations is achieved by reducing number of coefficients from 26 to 13 in case of children's speech. The plots reinforce the observations that children's speech will have shift in the formant values towards the higher frequency range and relatively more fluctuations in the spectrum due to higher pitch values.

From the above observations, we can expect that in case of children's speech recognition under mismatched condition, the performance to be significantly poor compared to matched condition of adults' speech recognition. The observations also suggest that higher band (3400-8000 Hz) needs to be preserved and reducing the fluctuations in the smoothed spectrum for the case of children's speech to get good performance. The different approaches for reducing the fluctuations present in the short term spectrum of children's speech and hence improving performance of children's speech recognition under mismatched condition was earlier explored [1].

1. Introduction

The present work focuses on extending the bandwidth of speech, which is more relevant for the children's speech, and accordingly trying to improve the children's speech recognition under mismatched condition.

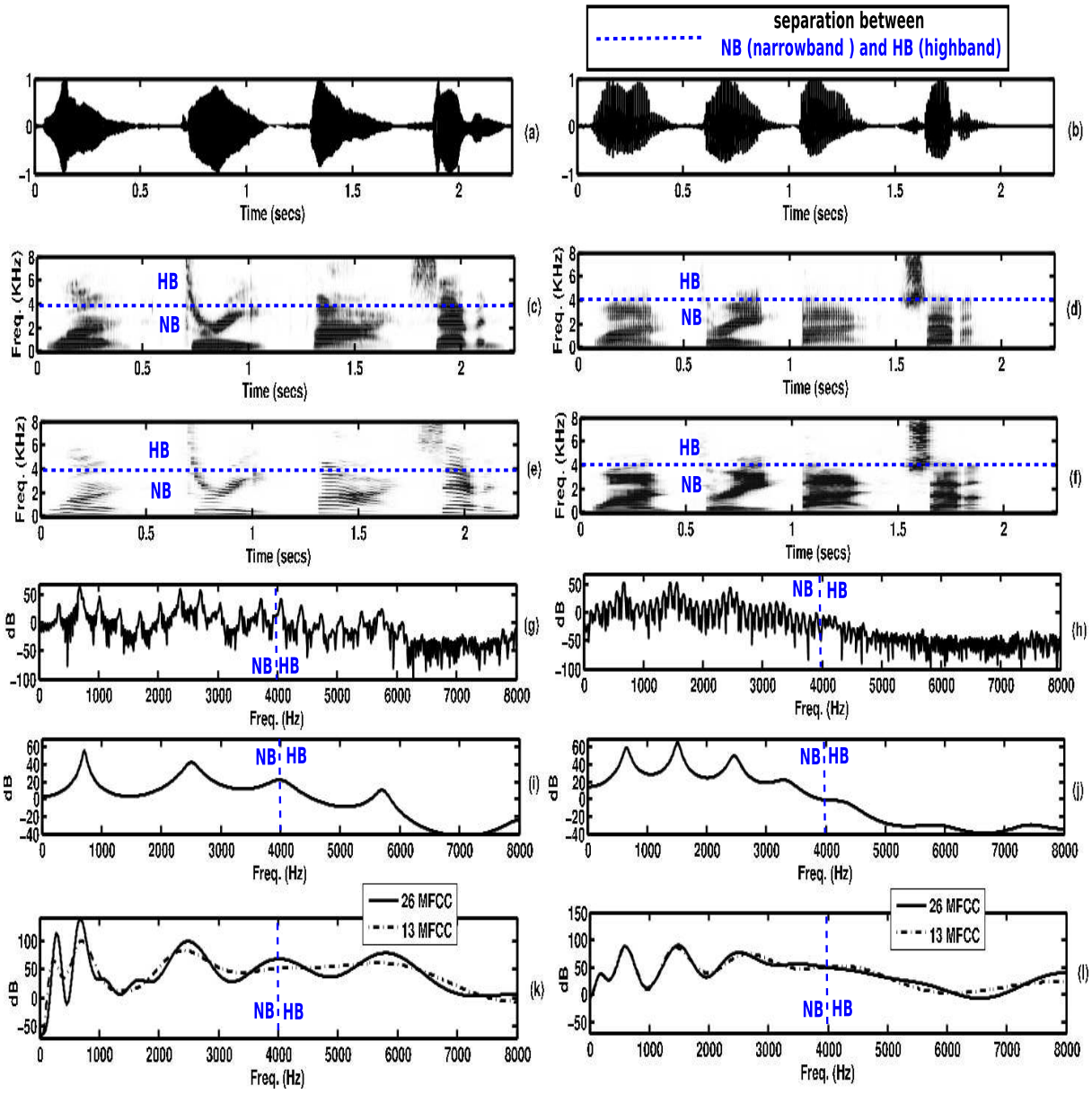

Figure 1.1: Differences in the nature of speech signals of a child (7 years old) and adult male for the English utterance 'one three five seven'. The first column represents the various plots related to child case and the second column for the adults' case. (a) & (b) waveforms, (c) & (d) wideband spectrograms, (e) & (f) Narrowband spectrogram, (g) & (h) STFT magnitude spectrum, (i) & (j) LPC spectrum, (k) & (l) smoothed magnitude spectrum from mel cepstrum. In each of the columns, the bottom three plots correspond to voiced frames for the same vowel /aa/ extracted from the speech signal.

1.2 Importance of bandwidth for speech recognition

The importance of bandwidth is demonstrated for both human recognition task and automatic speech recognition. The human recognition task involves subjective studies to identify the naturalness and intelligibility of speech. The subjective studies based on auditory perception have shown that reducing the speech bandwidth decreases the perceived speech quality. A progressive decrease in naturalness is observed when the upper cut-off frequency was decreased from about 11 kHz down to about 3.5 kHz [2]. Similarly, a marked degradation of naturalness is observed when the lower cut-off frequency increased from 123-208 Hz [2]. Apart from cut-off frequency value, the spectral balance between low and high frequencies is also observed to be an important factor for perceived naturalness. By that it means, the lack of naturalness caused by a high lower cut-off frequency cannot be compensated by changing the upper cut-off frequency. Similarly, the lack of naturalness caused by a low upper cut-off frequency cannot be compensated by changing the lower cut-off frequency. In another study related to extending the bandwidth, it is reported that extending the bandwidth from 300-3400 Hz to 50-3400 Hz is more beneficial for listener preference than extension to the range 300-7000 Hz. Also, using the full wideband range 50-7000 Hz gives the highest scores for perception.

Figure 1.2 (adapted from [2]) illustrates the effect of the audio bandwidth on the quality and intelligibility of speech. Figure 1.2(a) illustrates that high MOS score (representing more naturalness) is obtained for cases having low values for lower cut-off frequency and high values for higher cut-off frequency. The best MOS is for the case of 0 Hz for lower cut-off and 11 kHz for higher cut-off frequency. Thus from naturalness point of view, extended bandwidth in both lower and higher frequency directions from the narrowband (300-3400 Hz) bandwidth is desirable.

Figure 1.2(b) illustrates the importance of bandwidth for intelligibility. The low-pass filter cut-off frequency is shown as a curve. If the the cut-off frequency is about 0.1 kHz, then the intelligibility of speech is almost 0 % indicating that almost all the components in the speech are significantly attenuated. When the low-pass filter cut-off frequency reaches to about 5 kHz,

the intelligibility of speech increases to about 95 %. The high-pass filter cut-off frequency is also shown as a curve. If the the cut-off frequency is about 0.1 kHz, then the intelligibility of speech is almost 95 % indicating that almost all the components in the speech are passed. When the high-pass filter cut-off frequency reaches to about 5 kHz, the intelligibility of speech decreases to about 0 % indicating significant attenuation of almost all the speech components. Thus starting from 0 Hz, till about 5 kHz all the frequency components are important for speech recognition in case of adults' speech.

The human recognition studies for naturalness and intelligibility therefore infer that larger bandwidth of speech is desirable and also the higher scores depend on the larger bandwidth. Accordingly, advocate that the primary factor to be improved is bandwidth extension. The best spectral balance can probably be achieved if the bandwidth of telephone speech could be extended both below and above the conventional telephone band termed more commonly as narrowband (300-3400 Hz).

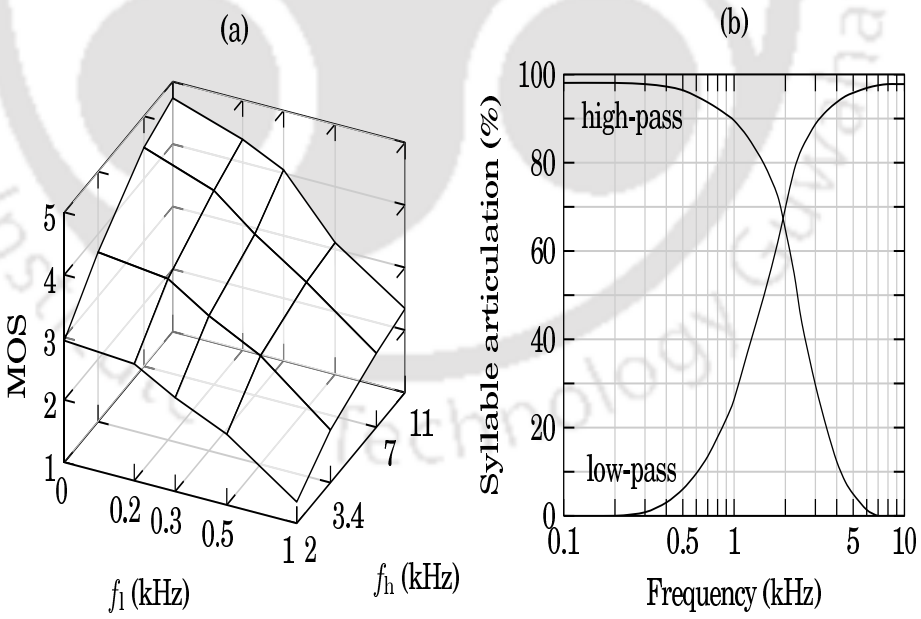

Figure 1.2: The effect of audio bandwidth on the quality and intelligibility of speech. (a) The speech quality measured using the subjective MOS scale is shown for different bandwidth limitations. The passband is determined by the lower (f_l) and upper (f_h) cut-off frequency of the bandpass filter. Data from Krebber (1995, figure 5.6) [3]. (b) The syllable articulation of lowpass and highpass filtered signals is shown as the function of the cut-off frequency. The syllable articulation is the percentage of correctly identified meaningless syllables. Data from French and Steinberg (1947, figure 12) [4]. Figure adopted from [2,5]

The effect of bandwidth on automatic speech recognition task is reported in the context of children's speech recognition [6], where both training and testing are done using children's speech. Figure 1.3 shows the word error rate (WER) expressed in percentage for children's speech recognition. In Figure 1.3(a), the studies on PSR database are reported [7,8]. The PSR word set is a list of 1000 words judged suitable for reading by 5 to 7 year-old children. The corpus consists of three subsets as follows: "PSR-Child(A)" contains recordings of five children judged to have good pronunciation, "PSR-Child(B)" contains recordings of seven children with varying levels of pronunciation proficiency and "PSR-Adult" consists of recordings of 10 adults aged between 19 and 34. The sampling frequency is 20 kHz for PSR- Child(A) and (B) and 16 kHz for PSR-Adult. If the bandwidth chosen is about 1 kHz, then WER is about 95% and decreases to about 40% when the bandwidth is 8 kHz. Also, the WER for adults' case falls below 20% when bandwidth is 4 kHz, but the WER is still very high for children. The same observation can be made for another children's speech recognition task using PF-STAR database given in Figure 1.3(b) [9]. The PF-STAR corpus contains 14 h of recordings from 158 British children (52% male). The training, evaluation, and test sets contain 7 h 29 min, 53 min and 5 h 49 min of recordings, from 86, 12, and 60 children, respectively. The test and evaluation sets contain only speech from 6 to 11 year olds, balanced by age. The data are sampled at 22.05 kHz. Both these studies infer that for automatic speech recognition task also, especially, children's speech recognition, larger bandwidth is necessary.

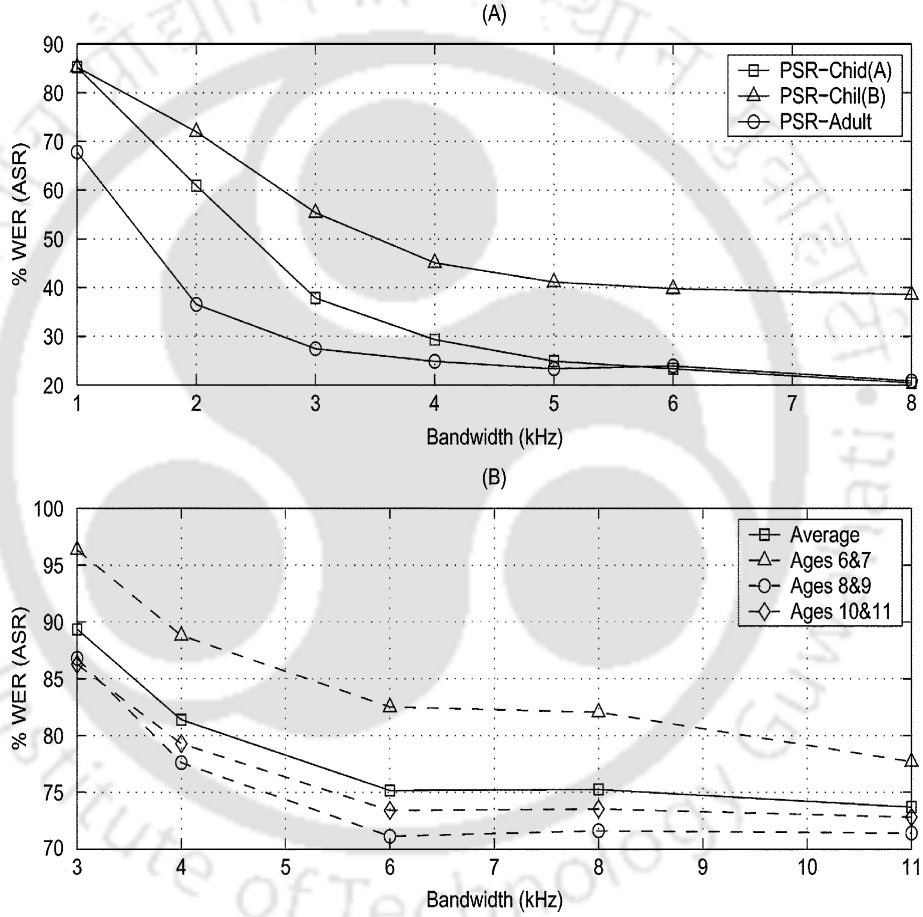

Figure 1.3: Effect of bandwidth on ASR performance on children's and adults' speech. (A) PSR corpus and (B) PF-STAR children's corpus. The figure is adapted from [6] ©2007 IEEE.

The instances taken from the literature and described above for the case of human and automatic speech recognition demonstrate the significance of larger bandwidth for speech recognition. Further, in case of children's speech, preservation of higher bandwidth is critical to obtain good performance. In case of narrowband speech, since the spectral energies are attenuated beyond 3400 Hz, the speech recognition performance for children's speech recognition will be significantly poor as compared to that of adults' speech case. Accordingly, we need methods for reconstructing the spectral energy information in the higher band from 3400 to 8000 Hz. The present work aims to develop new methods for the same which are collectively termed as artificial bandwidth extension (ABWE) methods.

1.3 Artificial bandwidth extension for speech recognition

From the literature, unless specified, the term ABWE refers to the process of extending the bandwidth of telephone speech (300-3400 Hz). Most of the ABWE methods exploit or based on the classical source-filter model given in Figure 1.4. The vocal tract system is represented by a time varying filter that contains information about the vocal tract shape and its dynamics. The filter is excited by a signal which is either train of impulses or random noise, scaled by a suitable gain. If train of impulses are used as excitation, the resulting speech is termed as voiced speech. Alternatively, the unvoiced speech is due to the random noise excitation. The process of exciting the time varying vocal tract system by time varying excitation results in the non-stationary speech signal as the output. Typically, the adults' speech contains spectral energy spread up to 8000 Hz. As in the case of telephone, the speech signal is passed through a bandpass filter in the range 300-3400 Hz resulting in narrowband speech.

In case of narrowband speech, both the excitation source and vocal tract system information gets significantly attenuated. The ABWE deals with the process of reconstructing the missing source and system information as much as possible. The typical block diagram of a source-filter model based ABWE process is given in Figure 1.5. The short term speech analysis is performed on the narrowband speech signal to separate out the vocal tract and excitation source components. The type of speech analysis depends on the ABWE method that will be

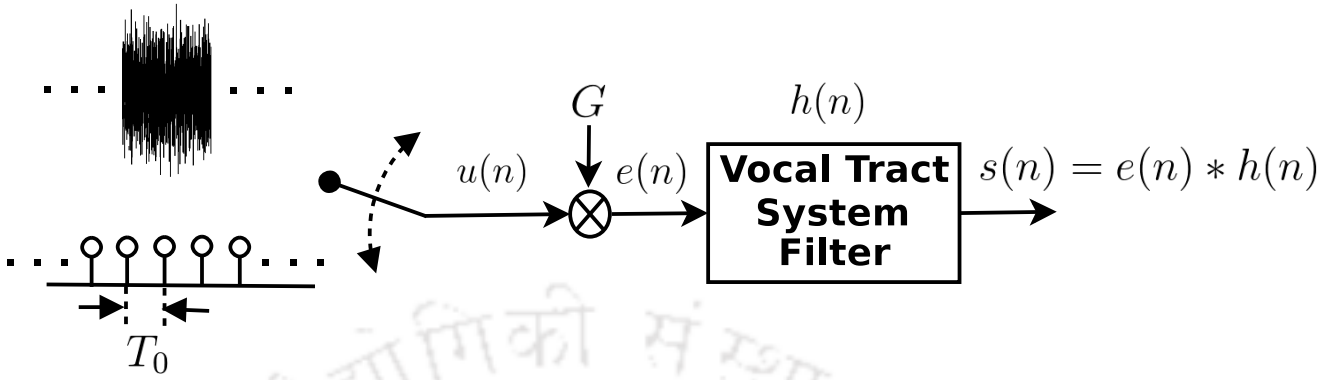

Figure 1.4: Two state speech production model depicting source-filter nature of approximation.

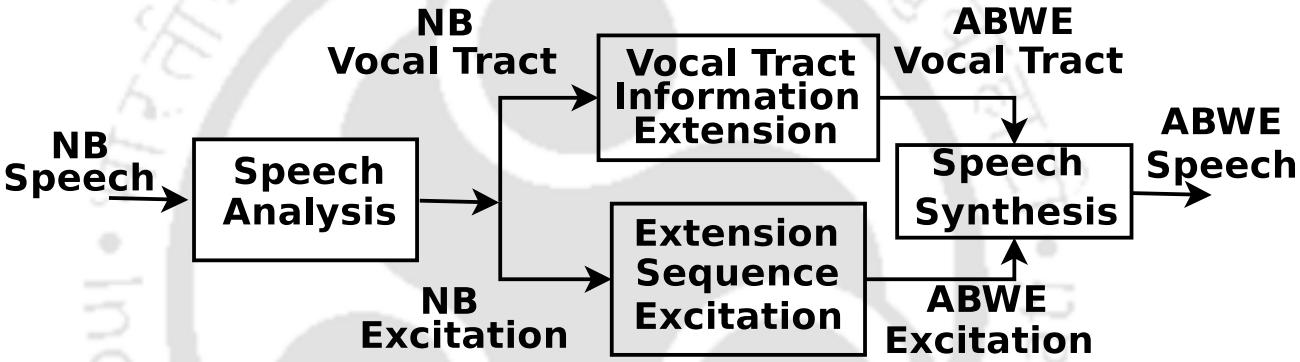

Figure 1.5: Generation of ABWE speech from NB speech.

employed. For instance, linear prediction (LP) analysis can be performed to extract the LP coefficients (LPC) representing the vocal tract information and LP residual representing the excitation source component.

The narrowband vocal tract information is subjected to a process of approximately reconstructing the missing vocal tract information in the range from 3400-8000 Hz. There are several methods in the literature for performing the same [10–14].The methods are usually based on the *a priori* information learnt during the training process. During the training process, the relation between the narrowband and wideband vocal tract information is learnt in terms of a mapping function. During testing, the incoming narrowband vocal tract information is mapped to the most relevant wideband vocal tract information using the mapping function. As a result, the artificially bandwidth (ABW) extended vocal tract information is obtained.

The narrowband excitation information is extended on either side of 300-3400 Hz to construct the excitation signal component in the frequency range less than 300 Hz and also more than 3400 Hz. Whether it is voiced or unvoiced speech, the excitation signal will have a flat spectrum. The process of reconstruction involves extending the flat spectrum beyond 300-3400 Hz. There are several methods in the literature like explicit signal generation, non-linear processing, modulation in time domain, spectral folding and so on for performing the same [15–21]. With the knowledge of source parameters these methods work to extend the bandwidth resulting ABW extended excitation signal.

Finally, the ABW extended vocal tract information is used as the filter information and excited by the ABW extended excitation signal to synthesize ABW extended speech signal. Since, both the vocal tract and excitation components are extended beyond the narrowband frequency range, the resulting speech is expected to provide more information for both human and automatic recognition task.

Most of the focus in the ABWE related work is for speech enhancement and speech coding. The main objective here is to provide improved quality and intelligibility for human perception. However, the improvement with respect to human perception may not directly correlate with improved performance in automatic speech processing tasks like speech recognition and speaker recognition [2, 4, 22–36]. Hence, there is a need for developing ABWE methods that provide improvement mainly in automatic speech processing tasks. Accordingly, they may not necessarily provide improved scores for human perception. Further, for children speech case, due to high non-stationary compared to adults' speech, both the vocal tract and excitations dynamics are very high resulting in significant mismatch. Accordingly, ABWE for children's speech needs a detailed exploration keeping in view of these dynamics. Hence the motivation for the present work.

To develop ABWE methods that are suitable for children's speech processing, understanding of the important differences between the adults and children is of paramount importance. Accordingly, the literature review needs to be done with respect to the works related to children's speech processing. After this, the existing methods for ABWE are to be reviewed to

understand how the current methods perform ABWE. Based on both these literature reviews, possible directions for the present work may be organized.

1.4 Organization of thesis

Chapter 2 provides a detailed review on the differences between children's speech acoustic-phonetic parameters compared to that of adults' speech. The different methods proposed in the literature for ABWE are also reviewed. Based on this, the organization of the present work is presented in the last section.

Chapter 3 describes a set of experimental studies to demonstrate the need for developing ABWE methods. These include development of baseline ASR system using TI-DIGITS database, implementing existing standard ABWE method and using it for ASR study. The studies also include comparing ABWE methods with vocal tract length normalization (VTLN) and truncation of coefficients methods.

Chapter 4 proposes a set of ABWE methods that are suitable for children's speech recognition under mismatched condition. They are collectively termed as methods using auxiliary information. The ABWE methods based on class-specific, age-specific and delta features are proposed in this chapter. A feature domain computationally efficient ABWE methods is proposed using mel frequency cepstral coefficients (MFCC).

Chapter 5 proposes ABWE methods using sparse representation framework. Initial version exploits the coupled dictionaries. The later versions are based on semi-couple dictionaries based and also using clustering.

Chapter 6 summarizes the different explorations, conclusions drawn from different explorations and possible directions for future work.

2

Literature Review

Contents

As discussed in the introduction chapter, the objective of this work is to develop methods for improving children's speech recognition under mismatched condition. The main causes of mismatches are due to shorter vocal tract length, and higher source dynamics. The earlier work attempted on improving children's speech recognition under mismatched condition by considering the effects of excitation source and minimizing them [1]. The focus of the current work is on exploring the effect of shorter vocal tract length and hence come up with methods to minimize its effects. As briefly outlined in the previous chapter, the short vocal tract length shifts the formants in case of children to a higher frequency range. Accordingly, methods are needed to approximate or reconstruct the higher band information from the given narrowband speech. These methods may be collectively termed as artificial bandwidth extension (ABWE) methods. To summarize, the goal is develop ABWE methods suitable for improving performance in case of children's speech recognition under mismatched condition.

To develop ABWE methods suitable for children's speech recognition, the understanding of the literature in the field of children's speech processing, especially, differences between children's and adults' speech is needed. Also, the existing methods for ABWE, even though meant for speech enhancement and coding applications, needs to be reviewed. Based on the understanding of both the fields, the directions for the current work can be laid out. Accordingly, this chapter is divided into four parts. The first part reviews all the works related children's speech processing. The second part reviews existing attempts for children's speech recognition under mismatched condition. The third part reviews the existing ABWE methods, starting from the earliest to the most recent ones. The fourth part discusses the requirements for bandwidth extension in case of children's speech and accordingly proposes directions for the work to develop methods for ABWE suitable for children's speech recognition.

The introduction to the children's speech recognition is given in Section 2.1. Section 2.2 highlights the differences between children's and adults' speech production systems. The acoustic mismatch differences in children's and adults' speech due to the differences in the speech production systems are described in Section 2.3. The differences in the linguistic correlates of children's and adults' speech are given in Section 2.4. Section 2.5 narrates the differences in

the short term features of children's and adults' speech. The different approaches for improving ASR Performance under mismatched condition existing in the literature are reviewed in Section 2.6. The need for ABWE for ASR under mismatched condition are hypothesized in Section 2.7. Section 2.8 provides review of existing attempts for ABWE. The special requirements of the ABWE suitable for children's speech recognition under mismatched condition are highlighted in Section 2.9. The final section (Section 2.10) describes the organization of the proposed work.

2.1 Introduction to children's speech processing

Most of the speech technologies developed in practice typically collect speech data from adult population. The reasons are several, mainly, it is expected that the intended target population for the developed technologies will be adult. Also, adult population is expected to be more cooperative compared to children. However, the scenario has changed where children are also interested in using speech technologies that may be available over hand held devices. The use of speech technologies developed using adult population by the children is demonstrated to be giving very poor performance [37–40]. There are two solutions to this, namely, either collect speech data separately for children population and rebuild the whole system using children's speech, or develop methods to improve the performance for adults' speech trained system itself. The former one is time and resource intensive and hence not recommended. The latter one is therefore the most preferred approach for making children to use speech technologies.

The later approaches are based on the fact that there is significant mismatch in the speech signal characteristics of children's speech compared to adult. The children's speech is highly non-stationary and exhibits variations in spectro-temporal patterns due to the shorter vocal tract length and higher source dynamics. To develop methods to reduce the mismatch, we need to understand the characteristics of children's speech and also differences with respect to adults' speech. The following sections describes them.

2.2 Children's and adults' speech production systems

The difference in the speech signal characteristics of children's and adults' speech stems from the fact that the two speech production systems and their associated dynamics are quite different. It will be interesting to get a feel about some of these differences.

A related work on the acoustic evidence for the development of speech explains the differences clearly [41]. Some of the important ones by considering a three years old child and an adult male are summarized as follows:

- The height of head is reported to be 18.8 cm for the child where as 23.6 cm for the adult.
- The length of the vocal tract for the child and the adult are found to be 10.4 cm and 16.9 cm, respectively.
- The lung vital capacity for the child and the adult are measured as 940 and 4450 cm^3, respective period.
- The rate of respiration in case of child and adult are observed to be about 24 and 12 breaths/min, respectively.
- The length of vocal folds is 0.45 cm for the child and 1.8 cm for the adult.
- The mass of vocal folds for the child and adult are reported to be 0.031 and 0.141 gm, respectively.
- The effective stiffness of the vocal folds are measured as 2.3×10^4 and 7.4×10^4, dyne/cm, respectively for the child and adult.
- The sub-glottal pressures are noted as 12 and 18 cm of H_2O, respectively for child and adult.
- The articulator repetition rate is observed to be 3 Hz for the child and 6 Hz for the adult.

The dimensions of head and vocal tract length will influence the resonance (formant) characteristics of the resulting speech. Since as mentioned above, the dimensions for children is

smaller compared to adult, the formant characteristics are expected to be significantly higher in the case of children. The lower lung capacity of children results in higher respiration rate and hence increases the non-stationary characteristics in case of children's speech. The vocal folds related measurements of children are significantly lower compared to that of adult. As a result, the vocal folds will vibrate at a much higher rate compared to the adults' case. Accordingly, the pitch period in case of children speech is expected to be much smaller compared to adult. The source dynamics therefore increases the non-stationary characteristics of children's speech. Even though, the stiffness of vocal folds of children is more, the faster vibration is maintained by the higher sub-glottal pressure. The speaking rate of the children is much lower compared to adult due to the lower articulation repetition rate.

2.3 Acoustic mismatch differences in children's and adults' speech

Several researchers have examined the differences in acoustic mismatch of children's and adults' speech. The age dependent changes in the formant frequencies and the fundamental frequency measurements of children speakers aged three to thirteen were reported in [42, 43]. Children have higher formant values for the same sound as compared to adult, and the formant values increase in a non-linear fashion [42]. The pitch values of children are relatively higher compared to adult causing large spacing between the pitch harmonics [42]. These high formant frequencies and pitch values are attributed to their inherent shorter vocal tract and vocal folds lengths, respectively. For instance, 5 year old children have been reported to have 50% higher value of formant frequencies than that of adult males [42]. In comparison to the presence of 3-4 formants for adults' speech, only 2-3 formants are present for children's speech within 0.3-3.4 kHz frequency range [40]. The higher formants of children's speech fall outside the narrow transmission bandwidth (3.4-4.0 kHz) of telephone channels resulting in the loss of spectral information in case of ASR of narrowband children's speech. This necessitates the requirement of artificial bandwidth extension methods for reconstructing the formants in the higher frequencies.

The phoneme durations and the average sentence durations have also been observed to be longer than that for adults which in turn reduces their speaking rate also [39, 41–43]. The average vowel durations of 5 year old children is reported to increase by 36% compared to those of 12 year old children [42]. On analyzing the consonant-vowel transition in case of adults and children's speech, it is noted that the children's speech have shorter transition duration and larger spectral difference between consonant and vowel in the consonant-vowel pair than those of the adults speech. Studies have found systematic decrease in the values of the mean and variance of the acoustic correlates such as formants, pitch and duration with age, with their values reaching adult ranges around 13 or 14 years [42].

The bandwidth of formant frequencies are observed to be dependent on radiation, glottal resistance, viscosity, heat conduction, and wall resistance [41]. The bandwidth contribution due to radiation (B_r) is directly proportional to the square of the formant frequency under consideration. The bandwidth contribution due to wall resistance (B_w) is inversely proportional to the square of the formant frequency under consideration. The bandwidth contribution due to both viscosity (B_v) and heat conduction (B_h) is observed to be directly proportional to the square root of the formant frequency under consideration. The bandwidth contribution due to glottal resistance (B_g) is directly proportional to the area of glottis. Among all these, the major factors that contribute to the bandwidth include B_g and B_r, where B_g contributes in a major way to the bandwidth of first formant and B_r contributes to the bandwidth of higher formants.

For children's case, the bandwidth contributions due to the effects of radiation are observed to be greater in comparison to that of adult. The reasons for the same is due to the higher formant frequency values for the children and also differences in vocal tract dimensions. The bandwidths of the first three formants due to radiation are 8, 71 and 197 Hz for children, and 3, 27 and 74 Hz for adult. significantly higher values can be observed for children.

The bandwidth contribution due to the effect of glottal resistance is observed to be more for children compared to the adults. The bandwidth of first formant due to glottal resistance is 66 Hz for children and 53 Hz for adults, showing higher value for children's case. The contributions

of other factors is observed to be same for both children and adults. Including all the factors, the bandwidth of the first three formants are 90, 132 and 246 Hz for children, and 72, 78 and 115 Hz for adult. By comparing different values for a given bandwidth, it can be agreed that major contributes for bandwidth differences among children and adults are radiation and glottal resistance.

To summarize, the acoustic mismatches between children and adults can be studied mainly by considering the changes in three factors, namely, higher formant values, higher pitch value and larger bandwidths in case of children as compared to that of adults. The different ways to minimize the effect of pitch changes was studied earlier [1]. To further increase the effectiveness, methods for minimizing the effect of mismatch in case of formants and their bandwidth is needed.

2.4 Linguistic correlates of children's and adults' speech

Apart from the acoustic mismatches, the linguistic mismatches between children and adult are also observed to be higher. The major causes of the linguistic differences are due to less control over the articulators, lesser vocabulary size and also mispronunciations in case of children.

Children's exhibit less precise control of the articulators especially at the age of 5-6 years. Sometimes they have not yet learnt how to articulate specific phonemes [44]. As a result, children's speech have many problems like dis-fluencies, false-starts and extraneous speech [45–47]. The frequency of occurrence of mispronunciations was noted in [47] to be almost twice as high for the 8-10 years old children than for the 11-14 years old children. Children have smaller vocabulary than adults and so, they use less words per utterance to convey the same message which correlates with their smaller lung capacity. The correct forms of certain words may not have been acquired fully by children, especially for those words that are exceptions to common rules. So, sometimes their sentences contain some spurious words which are not found in adults' case.

On exploring children read speech and spontaneous speech, similar trend was noted in their

linguistic variabilities with age in both cases. Children's spontaneous speech was also found to be less grammatical than adults' speech. However, the adult-level values were found to reach 1-2 years earlier for read speech [48]. Linguistic variability in children's speech reduces with age. Older children use simpler linguistic constructs and shorter utterances to convey the intended message. Dis-fluencies decrease with age and children reach adult-skill level at around 12-13 years of age (somewhat earlier for boys than girls) [48]. So, the ability of the children to use language efficiently to convey the message improves with their age.

2.5 Short-term features of children's and adults' speech

The differences in the vocal tract dimensions, glottal structure and associated dynamics results in changes in the formant values and their bandwidth, and pitch values. Accordingly, the speech signal nature is quite distinct in case of children's and adults' speech. In a generic sense, we can state that children's speech is highly non-stationary compared to that of adults' speech.

For any practical speech processing task like automatic speech recognition (ASR), short term processing is performed to extract relevant features for further modelling and testing. The significant difference in the performance of ASR system under mismatched condition can be attributed to the differences lying at the short term features level for the case of children's and adults' speech. There are several studies in the literature which have studied this and demonstrated that if steps are taken to reduce the variations present at the features level, then poor performance under mismatch can be significantly reduced.

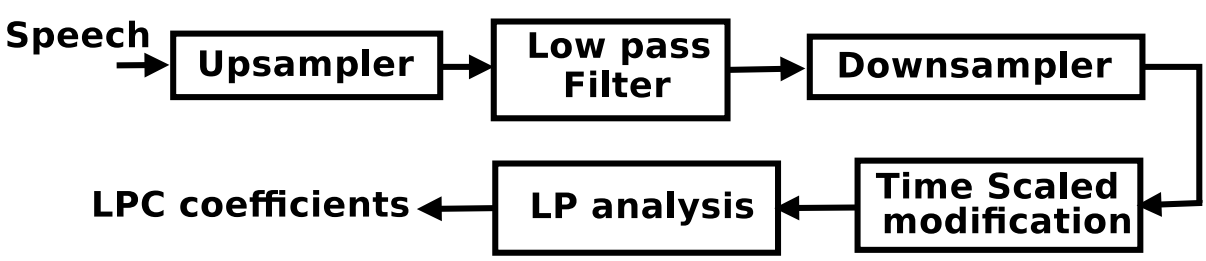

Figure 2.1: Steps involved in computation of LPC coefficients for high pitch valueed speech frame.

The effect of high pitch values on linear prediction (LP) analysis is reported in [49, 50].

It is stated that due to high pitch value as in the case of children's speech, the prediction is poor. The LP analysis of children's speech tries to model the spectral peaks corresponding to formants in a poor manner and is affected due to the high pitch values. Essentially, the LP analysis tries to model the pitch and its harmonics peaks rather than modelling the formant peaks. To minimize this effect, what is proposed is to first decrease the pitch period and then perform the LP analysis. The different steps involved are given in Figure. 2.1. This is done in the following way: up-sampling, low pass filtering to smooth out the fluctuations in the up-sampled signal, down-sampling, the ratio of up-sampling and down-sampling is done in such way that the overall pitch reduces by certain factor, say 1.5. This is achieved by increasing the pitch period by 1.5 using the above steps. However, the duration of the signal also increases by the same amount. The increase in the duration is compensated to the original value by time scaled modification. After this, the estimation of LP coefficients is made for the increased pitch case signal and is found to be better representing the vocal tract information compared to the original signal LP analysis. Thus this study demonstrates that, the LP coefficients are also affected in case of children's speech due to high pitch value. However, no attempt is made here to take care of increased formant values and their bandwidths.

The differences in the short term spectral measurements for the children's and adults' speech for the case of fricatives is reported in [51,52]. The measurements included spectral slope, spectral mean, spectral variance, spectral skewness and spectral kurtosis. These parameters are computed by constructing the short term spectrum of a 40 ms segment of speech using a 2048 point fast Fourier transform. By considering the resulting power spectra as the random distribution probabilities, from which the spectral moments are constructed. Measures of spectral slope were derived from the power spectra by a linear regression line fit to the relative amplitudes extracted from each analysis window. The spectral slope and mean are observed to be significantly lower for children. The spectral variance, skewness, and kurtosis are observed to be more for children. These studies infer that the short term spectral representation is the basic frequency domain representation of speech for extracting any parameter for further processing. The representation itself is significantly different in case of children compared to the adults

emphasizing the need for minimizing these variabilities to improve ASR performance under mismatched conditions.

A study on the use of pitch normalization for improving ASR performance under mismatched condition is reported in [53–55]. In this study, it is initially demonstrated that the variances in the higher dimension MFCCs is more in case of children compared to the adults. Also shown that the smoothed spectral envelope derived from the MFCCs show larger variations in case of children compared to adults. After this, several pitch normalization methods are used for reducing the variances present in the higher order MFCCs of Children's speech. The significant improvement in performance (reduction in WER) is observed for the pitch normalized case. Thus this study demonstrates that even though MFCC has smoothed the pitch variation, still the variances are large in case of children and hence the need further processing to improve performance. However, this study has focussed only on minimizing the effect of high pitch values in case of children's speech. The other variation due to higher formants and their bandwidths is not addressed.

The studies described above are the samples from the literature which show how the short term spectral features differ in case of children's and adults' speech cases. These changes can be attributed to both the changes in formants and their bandwidths and pitch values. Any improvement in ASR performance depends on how these variabilities are minimized. Therefore approaches are needed to minimize these effects at the feature level.

2.6 Approaches for improving ASR under mismatched condition

There are two main striking differences between the children's and adults' speech production systems. The short vocal tract length in case of children affects the formants and their bandwidths. The very thin vocal folds results in faster vibrations and hence very small pitch period or high pitch values. Both these aspects lead to high non-stationary in case of children's speech. Accordingly, poor ASR performance under mismatched condition. The following studies try to improve the ASR performance by minimizing the effect of these differences.

2.6.1 Vocal tract length normalization (VTLN) for improving ASR performance

A more commonly used approach for minimizing the differences between children's and adults' speech resulting due to the change in vocal tract dimensions is vocal tract length normalization (VTLN) [56–69]. VTLN aims to compensate for the fact that speakers have vocal tracts of different sizes. VTLN can be implemented by warping the frequency axis in the filterbank analysis. The warping factor α scales the distances of the filters in the mel filterbank. The optimal warping factor α is obtained by measuring the mean cepstral distance between adult and children's vowels. This warping is observed to be age dependent, 0.7 for 5 year old male child with acoustic models trained on male adults. For adult, $\alpha = 1$. The same warping can be used for all phones. However, phone dependent warping is observed to be further reducing spectral mismatch [40].

VTLN is a speaker normalization method in which the inter-speaker acoustic variability due to varying vocal tract lengths i.e., the mismatch due to difference in the formant frequencies among speakers is reduced by warping the frequency axis of the speech spectrum of each speaker [70,71]. For warping the frequency axis of the utterances during computation of MFCC features, the piece-wise linear frequency warping of filterbank, as supported in the HTK [72], may be used. The spacing and the width of the filters in the Mel filterbank are changed while maintaining the speech spectrum unchanged.

As the warping would lead to some filters being placed outside the analysis frequency range, to avoid the same a piece-wise linear warping function of the frequency axis of the Mel filterbank is employed [73]:

$$g_\alpha(f) = \begin{cases} \frac{1}{\alpha} f & 0 \le f \le f_c \\ \frac{1}{\alpha} f_c + \frac{f_{\max} - \frac{1}{\alpha} f_c}{f_{\max} - f_c}(f - f_c) & f_c < f \le f_{max} \end{cases} \tag{2.1}$$

where, $f_{\max}$ denotes the maximum signal bandwidth (4 kHz in this work) and f_c is an empirically chosen frequency of 3.4 kHz.

The optimal frequency warp factor for the test signal is estimated based on a maximum likelihood (ML) grid search over a possible range of warp factors given a current set of acoustic

models under the constraint of the first-pass transcription of the test signal. For instance, for doing ML grid search, each speech feature is warped by 13 different factors ranging from 0.88-1.12 in steps of 0.02. Given the various warped features, the optimal value $\hat{\alpha}$, by which the frequency axis of speech spectrum is warped, is estimated as:

$$\hat{\alpha} = \arg\max_{\alpha} P(\mathbf{x}_i^{\alpha}|\lambda, W_i) \tag{2.2}$$

where, $\mathbf{x}_i^{\alpha}$ represents the warped feature for the ith utterance with frequency axis of speech spectrum scaled by factor α. λ represents the HMM based speech recognition model and W_i is the transcription of the ith utterance. W_i is determined by the first recognition pass using the unwarped feature set. Ideally, the effect of using an optimal scaling factor selected in this way for each utterance is that of normalizing the test speech data with respect to the average vocal tract length of the training population of the recognition model set λ, thus reducing the inter-speaker acoustic variability between the training and the test data.

2.6.2 Model adaptation techniques for improving ASR performance

The model adaptation techniques compute a set of transformations for the means and/or the variances of the models or for the features that are used to reduce the mismatch between an initial model set and the adaptation data. The ML estimates of all transformation matrices for adaptation are obtained by solving a maximization problem for a standard auxiliary function using the expectation-maximization (EM) technique on adaptation data. The standard auxiliary function used to estimate the transforms is:

$$Q(M, \hat{M}) = -\frac{1}{2}\sum_{r=1}^{R}\sum_{m_r=1}^{M_r}\sum_{t=1}^{T} L_{m_r}(t)[\mathbf{K}^{(m)} + \log(|\hat{\mathbf{\Sigma}}_{m_r}|) + (\mathbf{o}(t) - \hat{\boldsymbol{\mu}}_{m_r})^T\hat{\mathbf{\Sigma}}_{m_r}^{-1}(\mathbf{o}(t) - \hat{\boldsymbol{\mu}}_{m_r})] \tag{2.3}$$

where, M represents the current recognition model set, $\hat{M}$ represents the adapted model set, T is the number of observations, m_r denotes a mixture component, $\mathbf{O}$ represents the sequence of d-dimensional observations, $\mathbf{o}(t)$ denotes the observation at time t. $\boldsymbol{\mu}_{m_r}$ and $\mathbf{\Sigma}_{m_r}$ represents the mean vector and covariance matrix for the mixture component m_r and $\mathbf{K}^{(m)}$ subsumes all constants. $L_{m_r}(t)$ represents the occupancy probability for the mixture component m_r at time

t and is defined as,

$$L_{m_r}(t) = p(q_{m_r}(t)|M, \mathbf{O}_T) \tag{2.4}$$

where, $q_{m_r}(t)$ represents the Gaussian component m_r at time t, and $\mathbf{O}_T = \{\mathbf{o}(1), ..., \mathbf{o}(T)\}$ represents the adaptation data.

2.6.3 Maximum likelihood linear regression (MLLR)

In MLLR [74] model adaptation technique, a set of linear transformations for the mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ parameters of the Gaussian distributions $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ in HMM are estimated. The effect of these transformations is to shift the component means and alter the variances in the initial system so that each state in the HMM is more likely to generate the adaptation data.

2.6.3.1 MLLR-MEAN

The adaptation method in which the linear transformations of only the means of the Gaussian distributions of the models are learnt using MLLR is referred to as 'MLLR-MEAN'. The transformation matrix used to estimate the adapted mean $\hat{\mu}$ is given by

$$\hat{\boldsymbol{\mu}} = \mathbf{W}\boldsymbol{\xi} \tag{2.5}$$

where, $\boldsymbol{\mu}$ represents a $d \times 1$ mean vector, $\mathbf{W}$ represents the $d \times (d+1)$ transformation matrix (where, d is the dimensionality of the data) and $\boldsymbol{\xi}$ represents the extended mean vector.

$$\boldsymbol{\xi} = [\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ...\boldsymbol{\mu}_d]^T \tag{2.6}$$

where, ω represents a bias offset which is kept as 1 (default value within HTK) in our work. Hence, $\mathbf{W}$ can be decomposed into

$$\mathbf{W} = [\mathbf{b}, \mathbf{A}] \tag{2.7}$$

where, $\mathbf{A}$ represents an $d \times d$ transformation matrix and $\mathbf{b}$ represents a $d \times 1$ bias vector.

2.6.3.2 MLLR-COV

The adaptation method in which the linear transformations are applied only to the variances of the models is referred to as 'MLLR-COV'. The transformation of the covariance matrix $\mathbf{\Sigma}$ is of the form

$$\widehat{\mathbf{\Sigma}} = \mathbf{H}\mathbf{\Sigma}\mathbf{H}^T \tag{2.8}$$

where, $\mathbf{H}$ represents the $d \times d$ transformation matrix.

This form of transformation can also be efficiently implemented as a transformation of the means and the features using the relation:

$$\mathcal{N}(\mathbf{o}; \boldsymbol{\mu}, \mathbf{H}\mathbf{\Sigma}\mathbf{H}^{\mathbf{T}}) = \frac{1}{|\mathbf{H}|}\,\mathcal{N}(\mathbf{H}^{-1}\mathbf{o}; \mathbf{H}^{-1}\boldsymbol{\mu}, \mathbf{\Sigma}) = |\mathbf{A}|\,\mathcal{N}(\mathbf{A}\mathbf{o}; \mathbf{A}\boldsymbol{\mu}, \mathbf{\Sigma}) \tag{2.9}$$

where, $\mathbf{A} = \mathbf{H}^{-1}$. Using this form it is possible to estimate and efficiently apply full transformations.

2.6.3.3 Constrained MLLR (CMLLR)

In this technique, a set of linear transformations for the features are estimated so as to modify the feature vectors such that their likelihood increase with respect to the given model.

In [75], it is shown that mean $\boldsymbol{\mu}$ and variance $\mathbf{\Sigma}$ of a Gaussian density $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma})$ associated with an HMM state can be adapted by means of an affine transformation, estimated in the maximum likelihood framework, in the following way:

$$\hat{\boldsymbol{\mu}} = \tilde{\mathbf{A}}\boldsymbol{\mu} + \tilde{\mathbf{b}}, \hat{\mathbf{\Sigma}} = \tilde{\mathbf{A}}\mathbf{\Sigma}\tilde{\mathbf{A}}^T \tag{2.10}$$

where, $\tilde{\mathbf{A}}$ and $\mathbf{b}$ represent the matrix and the offset vector of the so called constrained model-space transformation [75]. The term constrained denotes that the same matrix is applied to transform mean and variance. When a single transformation is used for adapting all the Gaussian densities in the recognition system, CMLLR adaptation can be implemented by transforming acoustic observations [75] using the following identity:

$$\mathcal{N}(\mathbf{A}\mathbf{x} + \mathbf{b}; \boldsymbol{\mu}, \mathbf{\Sigma}) = |\tilde{\mathbf{A}}|\,\mathcal{N}(\mathbf{x}; \tilde{\mathbf{A}}(\boldsymbol{\mu} - \mathbf{b}), \tilde{\mathbf{A}}\mathbf{\Sigma}\tilde{\mathbf{A}}^T) \tag{2.11}$$

In the feature-space transformation, to be applied to the feature vectors, represented by $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{b}}$ which are related to $\mathbf{A}$ and $\mathbf{b}$ by:

$$\tilde{\mathbf{A}} = \mathbf{A}^{-1}, \tilde{\mathbf{b}} = -\mathbf{A}^{-1}\mathbf{b} \tag{2.12}$$

Thus, the transformation matrix used to give a new estimate of the adapted observation is given by

$$\hat{\mathbf{o}} = \mathbf{W}_o\boldsymbol{\zeta} \tag{2.13}$$

where, $\mathbf{o}$ represents a $d \times 1$ observation vector, $\mathbf{W}_o$ represents the $d \times (d+1)$ transformation matrix (where, d is the dimensionality of the data) and $\boldsymbol{\zeta}$ represents the extended observation vector.

$$\boldsymbol{\zeta} = [\omega, \mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_d]^T \tag{2.14}$$

where, ω represents a bias offset which is kept as 1 (default value within HTK). Hence, $\mathbf{W}$ can be decomposed into

$$\mathbf{W} = [\mathbf{b}, \mathbf{A}] \tag{2.15}$$

where, $\mathbf{A}$ represents an $d \times d$ transformation matrix and $\mathbf{b}$ represents a $d \times 1$ bias vector. Since, multiple CMLLR transforms may be used it is important to include the Jacobian $|\mathbf{A}|$ in the likelihood calculation.

$$L(\mathbf{o}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{A}, \mathbf{b}) = |\mathbf{A}|\, \mathcal{N}(\mathbf{A}\mathbf{o} + \mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2.16}$$

2.6.4 Minimizing pitch mismatch for Improving ASR Performance

In this work [1], the acoustic mismatch due to pitch differences between the adults' and children's speech for children's ASR on adults' speech trained models are studied. It is found that apart from the formant frequencies, the pitch is the other major source of acoustic mismatch between the adults' and children's speech. The increase in the pitch of the signals is found to significantly increase the dynamic range and in turn the variances of the higher order coefficients of MFCC (C_0-C_{12}) features.

A pitch normalization algorithm is proposed which modifies the mel filterbank during MFCC

test feature extraction based on the average pitch of the test signal for children's ASR on adults' speech trained models. Also, a mel cepstral truncation based method is proposed for reducing the pitch mismatch between the training and the test data. The proposed algorithm automatically selects the appropriate length of the base MFCC features for each test signal without prior knowledge about the speaker of the test utterance. Significant improvements are obtained in the children's speech recognition performances using the proposed algorithms on the adults' speech trained models.

2.7 Need for ABWE for ASR under mismatched condition

From the review presented in the previous section, we can state that there are broadly three approaches by which the ASR performance can be improved under mismatched condition. The first and the mostly used approach is based on VTLN. This approach is a feature level approach for minimizing the differences between children's and adults' speech. In case of children's speech, the formants and their bandwidth shift towards higher frequencies. Accordingly, using the Mel filterbank with conventional spacing for the filters as in the adults' case is not appropriate. The frequency scale is therefore warped using a suitable warping factor to match for the shift in the formants for children. As a result, significant improvement in the performance is achieved. However, it is to be noted that VTLN is effective only within the given band of speech. In case of narrowband speech, VTLN will attempt to minimize the mismatch between children's and adults' speech within the given narrowband. However, the significant information about the formants and bandwidth extend much beyond 3.4 kHz for the case of children. In case of adult, most of the formants information is present within 3.4 kHz. If we want to improve the ASR performance for children's case, then some ways of first constructing the missing formants and their bandwidth in the higher frequency range is needed. This reconstruction of higher information may provide better match and hence improved performance. On top of extended bandwidth, VTLN may further improve the performance.

The model adaptation techniques essentially adapt the adults' models to suit more for children's speech, by estimating the mean and variances of the GMMs of HMM. Due to this

some improvement in ASR performance can be achieved. In the narrowband case, the children's speech used for adaptation still has only partial information. As a result the amount of gain achieved may not be significant.

The attempts to reduce the pitch mismatch will address only the effects due to source and do not focus on formants and their bandwidth that are present much beyond 3.4 kHz. Of course, the mismatch is significant for the pitch case itself. As a result any method to reduce the pitch mismatch also provides significant performance improvement.

To summarize, for the children's speech the pitch values are higher compared to adult. The effect due to pitch mismatch are already addressed in the literature as described above for improving ASR performance. The second difference in the children's speech is the shift of formants to the higher frequency range. This is addressed by using VTLN. In case of narrowband speech, there is yet another difference, that is missing higher formants and their bandwidth information beyond 3.4 kHz. If we want to achieve good improvement in ASR performance for narrowband case, then first we need to have a method for reconstructing the missing information beyond 3.4 kHz. When such extended bandwidth speech is used, the mismatch may significantly reduce. These attempts in the proposed work are collectively termed as ABWE methods.

2.8 Artificial bandwidth extension: A Review

This section reviews techniques proposed for the artificial bandwidth extension (ABWE) of narrowband (telephone) speech. The section begins with the motivation for ABWE and defines the frequency bands relevant for ABWE. Signal processing techniques utilized in ABWE are then described in detail.

2.8.1 Motivation for ABWE

The subjective studies have shown that the speech bandwidth significantly affects the perceived quality of speech, and wideband (0-8 kHz) speech is consistently preferred over narrowband (0-3.4 kHz) speech [36, 76–79]. Furthermore, the intelligibility of wideband speech

is found to be higher than that of narrowband speech [4]. The evaluations have shown that the intelligibility of whole sentences in narrowband speech is as high as about 99%, but the intelligibility of meaningless syllables is only about 90% [80–82]. Improving the intelligibility makes communication more comfortable and reduces the listening effort [80]. Thus, increasing the acoustic bandwidth of narrowband speech is clearly beneficial.

Wideband telephone calls require the network to support wideband speech transmission, and users need to have terminal devices designed for wideband signals at 16 kHz sampling rate [83]. During the transition period, both narrowband and wideband calls will occur depending on the capabilities of the network and the terminal devices, and users of wideband capable terminal devices will experience a difference between narrowband and wideband calls. To reduce the quality difference between narrowband and wideband speech, techniques have been developed to extend the bandwidth and thus to improve the quality and intelligibility of narrowband speech.

Bandwidth extension of telephone speech attempts to reconstruct missing frequency content outside the frequency range present in a narrowband speech signal. ABWE performs this task utilizing solely the information in the input speech signal. The input of a speech bandwidth extension system is usually assumed to be a speech signal produced by one talker at a time. Consequently, the characteristics of the speech production mechanism and speech signals can be exploited in the bandwidth extension task. Numerous experiments have shown that the perceived quality of the output of state of the art ABWE methods is higher than that of narrowband speech on the average.

2.8.2 Frequency bands

The purpose of ABWE is to reconstruct an approximation of wideband speech from narrowband speech. Digital narrowband speech is sampled at the rate of 8 kHz and therefore has a strict upper bound in the bandwidth at 4 kHz. Typically, narrowband speech is bandlimited approximately to the traditional telephone band (0.3-3.4 kHz) specified for PCM channels, but the exact passband varies. Wideband speech usually refers to the audio frequency range from

50 Hz to 7 kHz using a sampling rate of 16 kHz.

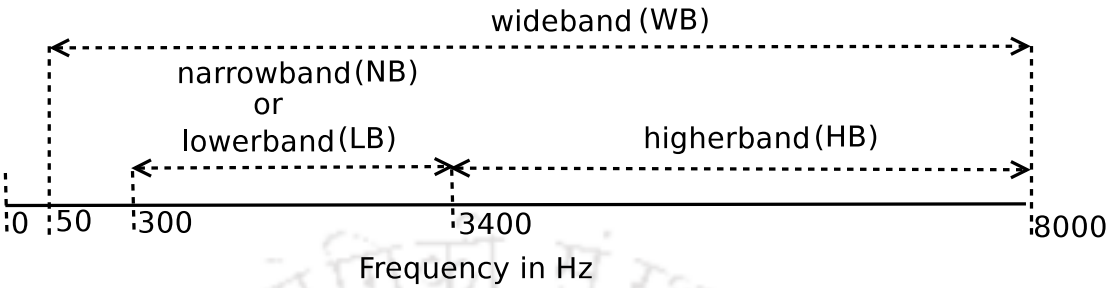

Figure 2.2: Frequency bands in the bandwidth extension of telephone speech.

The speech bandwidth can be extended to frequencies below or above the telephone band or both. The extension band above the telephone band typically ranges from 3.4 kHz up to 7 kHz and is called the higher band. The low-frequency extension band typically covers frequencies below 300 Hz, but the frequency limit may vary.

While a number of speech and audio codecs capable of transmitting wideband (50-7000 Hz), super-wideband (50-14,000 Hz), and even fullband (20-20,000 Hz) signals have been developed [84], narrowband speech transmission still prevails in cellular telephone networks and the transition to wideband speech is in progress. Consequently, the frequency ranges shown in figure above are currently the most relevant for the bandwidth extension of narrowband speech.

2.8.3 Correlation between frequency bands of speech

ABWE is based on the assumption of dependency between the contents of frequency bands in speech. The dependency between the frequency bands justifies the estimation of the missing frequency content from the bandlimited input. The assumption of dependency is reasonable because the entire speech spectrum is generated by the same physical and acoustic configuration of the speech production apparatus. As an example of a signal source characteristic affecting a wide range of frequencies, the spectral tilt of the glottal source signal causes an overall descending spectrum in both the telephone band and the higher band. However, the dependency between the frequency bands has turned out to be relatively weak [85, 86].

The correlations between the telephone band and both the lowband and the higher band

have been investigated with information theoretic measures. In [87], the authors have derived an upper bound for the quality of memoryless ABWE in terms of spectral distortion using an information theoretic approach. In [88], it is also examined the mutual information between the spectral shapes in the telephone band and in the higher band and concluded that memoryless ABWE cannot be expected to reconstruct the higher band spectrum accurately from the narrowband spectrum. The relationship between narrowband and higher band spectral envelopes is observed to be one to many and high quality reconstruction requires additional transmitted information [89]. It is also shown that including memory in the estimation technique improves the certainty of the higher band estimate [90].

In practice, ABWE cannot reconstruct the original wideband speech faithfully from a narrowband speech signal. Instead, the general aim of ABWE is to add energy to the extension bands in a perceptually reasonable way so that the perceived bandwidth is increased and the subjective speech quality is improved [86, 91, 92].

2.8.4 Speech bandwidth extension with side information

The higher band cannot be accurately estimated from the narrowband speech signal. Due to this, bandwidth extension techniques utilizing a small amount of additional transmitted information about the missing frequency range have also been developed [28, 30, 34, 93–96]. Auxiliary bit streams of about 130 bps-2 kbps are generally suggested to support the bandwidth extension. The additional information enables a more accurate regeneration of the missing frequencies than artificial bandwidth extension. As a result, fewer artifacts are produced and the overall quality is improved.

Utilizing side information with the existing narrowband systems also calls for a method to transmit the information to the receiving end. If interoperability with existing standard codecs is required, data transmission can be accomplished in three different ways [95]: by embedding the side information in the speech signal itself [97, 98], by modifying the encoded bit stream, or most efficiently by joint coding and data hiding within the encoder [25, 95, 96]. All these approaches maintain interoperability with the existing networks and codecs but may cause a

minor degradation in speech quality. The signal-domain approach is problematic because it is not sufficiently robust when coding with code excited linear prediction (CELP) codecs [25]. The third approach of including the data hiding within the encoding process provides the best results but requires codec dependent processing. The idea of bandwidth extension with a small amount of transmitted side information is also utilized in many standardized wideband speech and audio codecs, where the higher frequencies are estimated from the transmitted lower frequency content and additional side information parameters [97].

2.8.5 Different techniques for artificial bandwidth extension

The different techniques present in the literature for ABWE can be broadly grouped under the following categories.

2.8.5.1 General signal processing methods

One of the earliest studies on the artificial bandwidth extension of telephone speech was the investigation aimed at improving the acoustic quality of telephone contributions in broadcast programs [99]. Both lower band and higher band extension were considered challenging, and especially higher band extension was found to be extremely difficult using the analog techniques of the time. The first adaptive approach to speech bandwidth extension made a distinction between voiced and unvoiced speech, utilized frequency domain processing, and generated a high frequency extension that was scaled according to the short term spectral tilt of the band limited input signal [100, Section 1.3].

More advanced digital signal processing techniques have been proposed for artificial bandwidth extension since the early 1990s. Statistical recovery of the missing higher band using a combination of autoregressive filters excited by Gaussian noise was proposed [101] The use of codebooks for the spectral envelope extension was described [102,103]. Similar techniques form the bases of many of the more recent ABE methods.

In general, ABE methods utilize a combination of techniques from pattern recognition, statistical estimation, and speech synthesis [95]. Input speech is typically processed in short frames of about 10-30 ms. In some cases, the temporal characteristics of the signal are adjusted

with a finer resolution of about 2 ms. Successive frames may overlap, and output frames may be concatenated with overlap and add techniques.

Bandwidth extension can be implemented in the frequency domain or in the time domain. Frequency domain processing commonly computes the spectrum of the input frame using a fast Fourier transform (FFT), constructs the spectrum of a frame in the spectral domain, and applies the inverse FFT to produce the corresponding time domain signal. Time domain processing typically involves adaptive filters or a filterbank for the shaping of the extension band spectrum. Frequency domain processing has the benefit of accessing the spectral representation directly, whereas time domain processing may yield a lower overall delay and avoid potential degradations due to the limited accuracy of FFT computation especially on fixed point platforms.

2.8.5.2 Source-filter model based ABWE

A large part of the ABWE methods proposed in the literature are based on the source-filter model of speech production. The narrowband input signal is divided into an excitation signal and a filter representing the spectral envelope. The excitation and the filter are then extended separately to the extension band. The extension of the filter requires an estimation method that generates the filter parameters from a set of descriptive features calculated from the narrowband input. Bandwidth extended speech is generated by processing the extended excitation with the extended filter and combining the output with the original narrowband speech signal.

Extension of the excitation: The excitation signal contains the spectral fine structure of the extension band and usually has a relatively flat spectral envelope. Typically, the extension band excitation is generated from the narrowband excitation, which is often represented by the LP residual of the input signal. Several methods have been proposed for the extension of the excitation as described below.

Spectral folding: Spectral folding is a simple way to extend the excitation signal of a narrowband speech signal from the frequency range 0–4 kHz to 0–8 kHz. Spectral folding generates

a mirror image of the narrowband spectrum in the higher band [29]. It is equivalent to an up-sampling operation without an anti-aliasing low pass filter and makes use of the aliased spectrum in the extension band. Spectral folding can be implemented easily in the time domain by inserting a zero sample after each input sample or in the frequency domain by mirroring the FFT coefficients. Figure 2.3 illustrates the effect of spectral folding in the frequency domain.

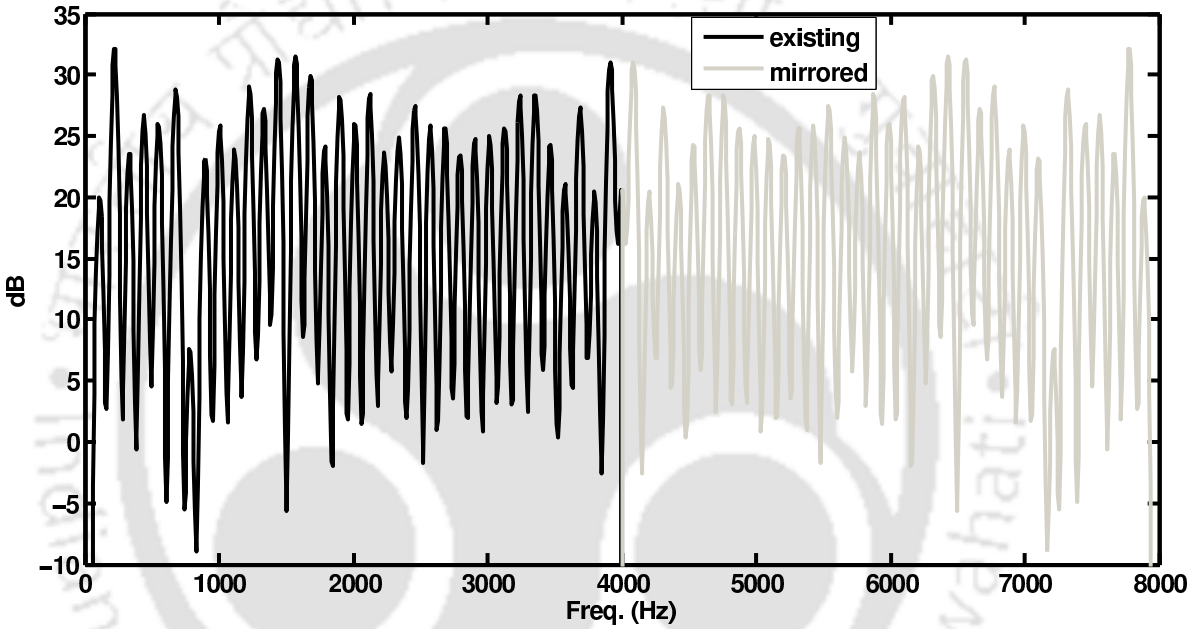

Figure 2.3: Spectral folding can be used to generate spectral content in the higher band. Zero samples are added between the signal samples in the time domain, which causes the spectrum to be mirrored to the higher band as shown.

Modulation techniques: A shifted copy of the up sampled baseband spectrum can be generated in the extension band with modulation at a fixed frequency. Filtering is necessary to prevent undesired overlapping of the translated spectrum and the original narrowband spectrum. The modulation frequency can be equal to 4 kHz, which corresponds to the spectral translation and is simple to implement in the time domain. Other modulation frequencies can also be used and allow, e.g., filling the gap in the spectrum at 4 kHz [80]. Fixed modulation breaks the harmonic structure because spectral peaks of voiced speech copied to the higher band are not located at integer multiples of the fundamental frequency. While the modulation is a time domain process as described, a similar effect can be achieved by frequency domain

processing. For example, the higher band excitation is constructed in the frequency domain by repeatedly copying a subband of the telephone band spectrum to the higher band [104, 105].

Pitch-adaptive modulation: The harmonic spectrum of voiced speech can be preserved by copying the narrowband spectrum to the extension band using an adaptive modulation frequency that is a multiple of the fundamental frequency (f_0) of speech [29]. Pitch adaptive modulation requires an accurate estimate of f_0. Pitch detection is a non-trivial task for which there are several basic methods and a number of enhancements to improve robustness and accuracy.

Sinusoidal synthesis: Sinusoidal synthesis generates the excitation signal of voiced speech as a sum of sine waves with frequencies equal to the multiples of the fundamental frequency of voiced speech. A similar result can also be achieved by generating harmonic peaks in the frequency domain. A mixed excitation with an adjustable level of harmonics for varying levels of voicing can be generated by means of randomized harmonic phases or additive random noise [106] [93, Section 2.4] [107]. Sinusoidal synthesis is especially suitable for the excitation extension to low frequencies below 300 Hz because the low band signal primarily consists of a small number of harmonics of f_0 and the ear is sensitive to the harmonic components in this frequency range. Sinusoidal synthesis calls for an accurate estimate of f_0. If ABWE is implemented in close connection with a speech decoder, the pitch period estimate utilized in many codecs can be exploited also in ABWE [29].

Nonlinear processing: The excitation signal can be extended by applying a non-linear function $f(x)$ to the narrowband excitation. Examples of non-linear functions used for bandwidth extension include $f(x) = x^2$ [20], $f(x) = x^3$ [108], and $f(x) = |x|$ [92, 109]. An adaptive non-linear function that adapts to the input amplitude is described in [22, Section 4.2.7]. Non-linear processing has the benefit of maintaining the harmonic character of the excitation, i.e., non-linear processing of a periodic signal generates a spectrum with spectral peaks at integer multiples of f_0. However, the energy level generated in the extension band by a non-linearity is difficult to control, and subsequent energy normalization is typically required [87, Section 3.2], [108]. Non-linear functions are often used for the low frequency extension because they

generate a harmonic spectrum, which is perceptually important at low frequencies.

Noise modulation: Noise modulation refers to generating the excitation by modulating a white noise signal by a temporal envelope. This technique can be motivated by the critical bandwidths of human perception: above 4 kHz the frequency resolution of the human ear is poor and pitch harmonics are not resolved individually, but pitch periodicity is present in the temporal envelope of voiced speech [30]. Thus, the harmonic spectrum does not need to be reproduced and the pitch periodicity can be reconstructed by the time domain modulation of a noise excitation. The temporal modulation envelope can be extracted from the time envelope of a sub-band of the input signal using, e.g., the frequency band of about 2-3 kHz [110], 2.5-3.5 kHz [111, 112], or 3-4 kHz [81].

Noise excitation: A noise signal has also been used as an excitation in combination with other techniques to avoid an highly periodic excitation at high frequencies [104] or to provide an excitation for unvoiced speech sounds [16, 106, 113]. For the extension from the wideband frequency range to the super-wideband range, a noise excitation may be sufficient, especially if the temporal envelope of the excitation is also adjusted as in the method described by [114].

Voice source modelling: A method was proposed for estimating the wideband voice source signal from the narrowband signal to extend the excitation of voiced speech [115]. The technique was found to be especially effective in the low frequency range. Furthermore, the bandwidth extension layer in the ITU-T G.729.1 codec [116] utilizes a lookup table of glottal pulse shapes to reconstruct the excitation for voiced speech [95].

2.8.5.3 Feature extraction

The spectral envelope in the extension band is estimated from the information available in the narrowband input signal. For this purpose, a set of features is calculated from each frame of the input signal. The aim of feature extraction is to compress the input frame into a small number of values that represent relevant characteristics of the signal for the envelope estimation task. The features should be selected so that they provide as much information about the missing frequency band as possible. At the same time, the number of features, i.e.,

the dimensionality of the feature vector, should be small to keep the computational complexity low.

A multitude of features have been proposed for bandwidth extension. Two instrumental measures can be used to quantify their suitability for the task [117]:

- The information theoretic quantity called mutual information describes the dependency between signals. The mutual information between a feature set and the quantity to be estimated indicates the feasibility of the estimation task.
- The separability quantifies the discriminative power of a feature set for the classification of speech frames into relevant categories.

Features can be classified into two categories: frequency domain and time domain features. Frequency domain features represent the characteristics of the spectrum and are typically computed from the FFT based magnitude spectrum of a frame. Time domain features are computed directly from the signal samples and represent the temporal characteristics of a signal frame. Many early ABWE approaches utilized only spectral envelope parameters of the narrowband input to estimate the spectral envelope parameters of the extension band. However, additional time domain and frequency domain features have been shown to be beneficial for the estimation [117].

The following list presents some typical examples of the features.

Subband energy levels: The overall spectral shape of the input signal can be represented by the amounts of energy in a small number of frequency bands [109, 118].

Autocorrelation coefficients: Alternatively, the spectral envelope can be represented by the first ten autocorrelation coefficients [80].

LPC filter coefficients: The coefficients of the all pole filter obtained by linear prediction can also be used to represent the spectral envelope [91]. LPC parameters can be converted to other representations to be used as input features, such as line-spectral frequencies (LSF) [107, 110, 119–122], mel-scaled LSFs [123], or linear prediction cepstral coefficients (LPCC) [124].

Cepstral coefficients: Another representation of the spectral envelope is provided by the

mel-frequency cepstral coefficients (MFCC) [125], which are commonly used as input features in automatic speech recognition [126]. MFCC features are used for ABWE [115, 127–129]. Alternatively, linear frequency cepstral coefficients (LFCC) can also be used [92].

2.9 ABWE for children's speech recognition

The earlier sections reviewed the developments in children's speech processing, approaches for improving ASR performance in case mismatch conditions and ABWE areas. The work in the children's speech processing area focussed first on understanding the differences between children's and adults' speech. As reviewed earlier, the differences are mainly due to the changes in the dimensions of the vocal tract and glottal excitation source. As a result, the formants shift to higher frequencies compared to adults' speech. The bandwidths associated with the formants will be more for children speech. The children's speech will have higher pitch values. Due to these, the acoustic features extracted from children's speech will be quite different from those of adults' speech. This can be observed through Figure 2.4 where the variances among the higher order coefficients is more between low and high pitch cases. As a result the models developed for adults' speech shows large error when tested with children's speech features.

The review of works related to ASR under mismatched condition revealed that the performance can be improved by one of the following approaches: VTLN, model adaptation and pitch mismatch minimization. Followed by this the need for ABWE for children's speech recognition in case of narrowband speech was also described.

With respect to the developments in ABWE, the focus is mainly on converting telephone quality speech to wideband or super wideband speech. Even though not stated explicitly, the implicit assumption an all ABWE work assumes adults' speech. This may be due to the large adult population user base for telecommunication. Also, the emphasis or focus of the ABWE work is to develop methods that can extend bandwidth of telephone speech which provides perceptually improved speech. As long as the ABWE speech gives perceptual comfort to the listener, no further requirements will be put forward.

However, the issues in children's speech recognition for improving ASR performance under

mismatched conditions are different. In case of children's speech, since most of the information extends beyond 3.4 kHz, a faithful reproduction of information from 3.4-8 kHz is necessary. The earlier attempts for improving children's speech recognition under mismatched condition focussed on minimizing the variabilities due to pitch [1,53–55]. Here, whether the given speech is either narrowband or wideband, the focus is mainly to minimize the fluctuations due to pitch. By doing the same, improvement in ASR performance is reported. In this the approaches based on pitch modification focus on minimizing the effects of pitch.

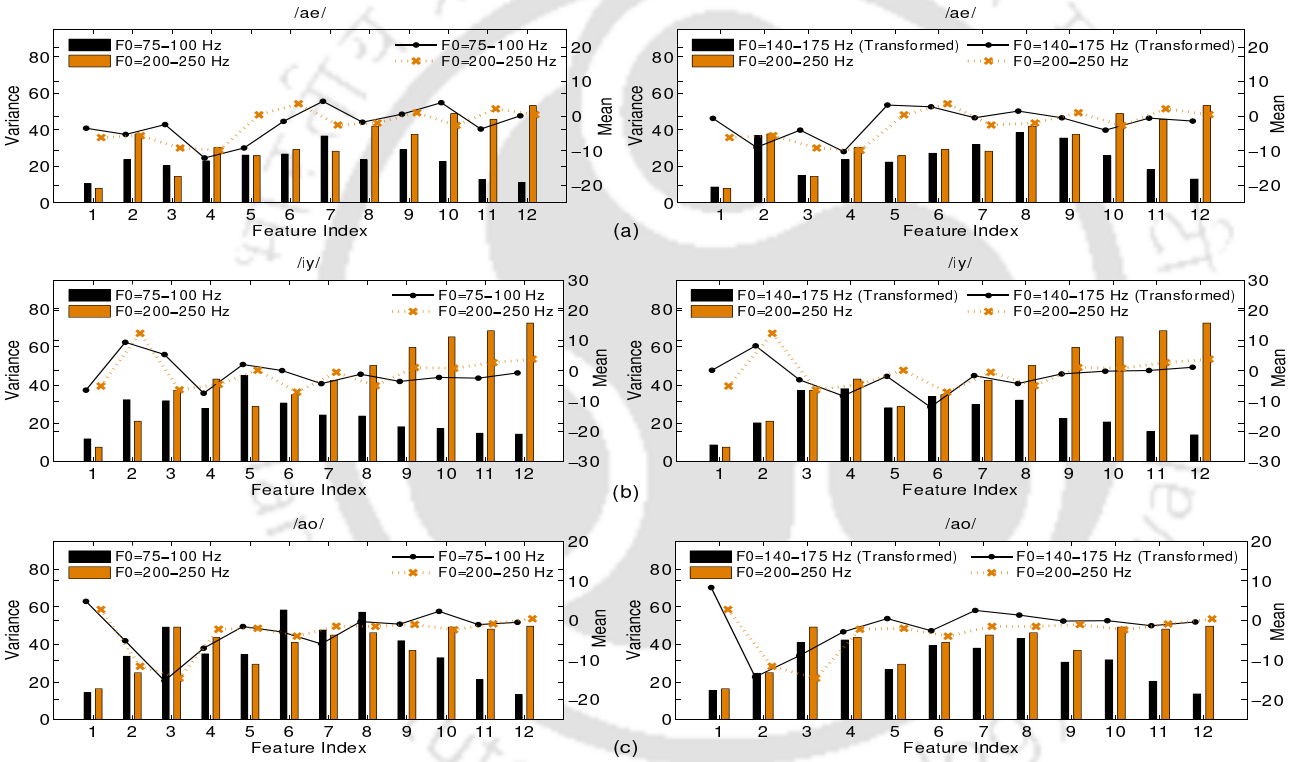

Figure 2.4: Plots showing mean along with variance (in bar) for MFCC (C_1-C_{12}) of signals of different pitch groups: 75-100 Hz and 200-250 Hz (left panel) and 200-250 Hz and its transformed to 140-175 Hz (right panel) for vowels (a) /ae/ (b) /iy/ (c) /ao/. [54] ©INTERSPEECH 2009

2.10 Organization of the Present Work

To minimize the mismatch due to increase in formants and their bandwidth of children's speech, the most commonly used methods are based on vocal tract length normalization (VTLN) [70, 72]. However, VTLN cannot provide complete solution as explained below. In Figure 2.5 the LP spectra representing the formants information for children's and adults' speech are shown for the same sound unit. The formant frequencies for the children shift to higher frequencies and their bandwidths increase. However, the most important observation is, VTLN can only take care of minimizing these variabilities only in the given band. That is, VTLN is effective for compensation of narrowband children's speech with respect narrowband adults' speech. Similarly, for wideband children's speech with wideband adults' speech. It fails to do the compensation for cross-band scenario. That is, the VTLN on narrowband children's speech is not effective for reconstructing the missing information that is present much beyond 3.4 kHz for child. As a result even though the first two formants of the child gets warped with adults' case, the higher formants information are still missing. In such scenarios what is suggested is to perform ABWE for given narrowband child speech and then compare with the wideband adults' speech. An additional VTLN on ABWE children's speech may further improve matching. Hence the motivation for the present work.

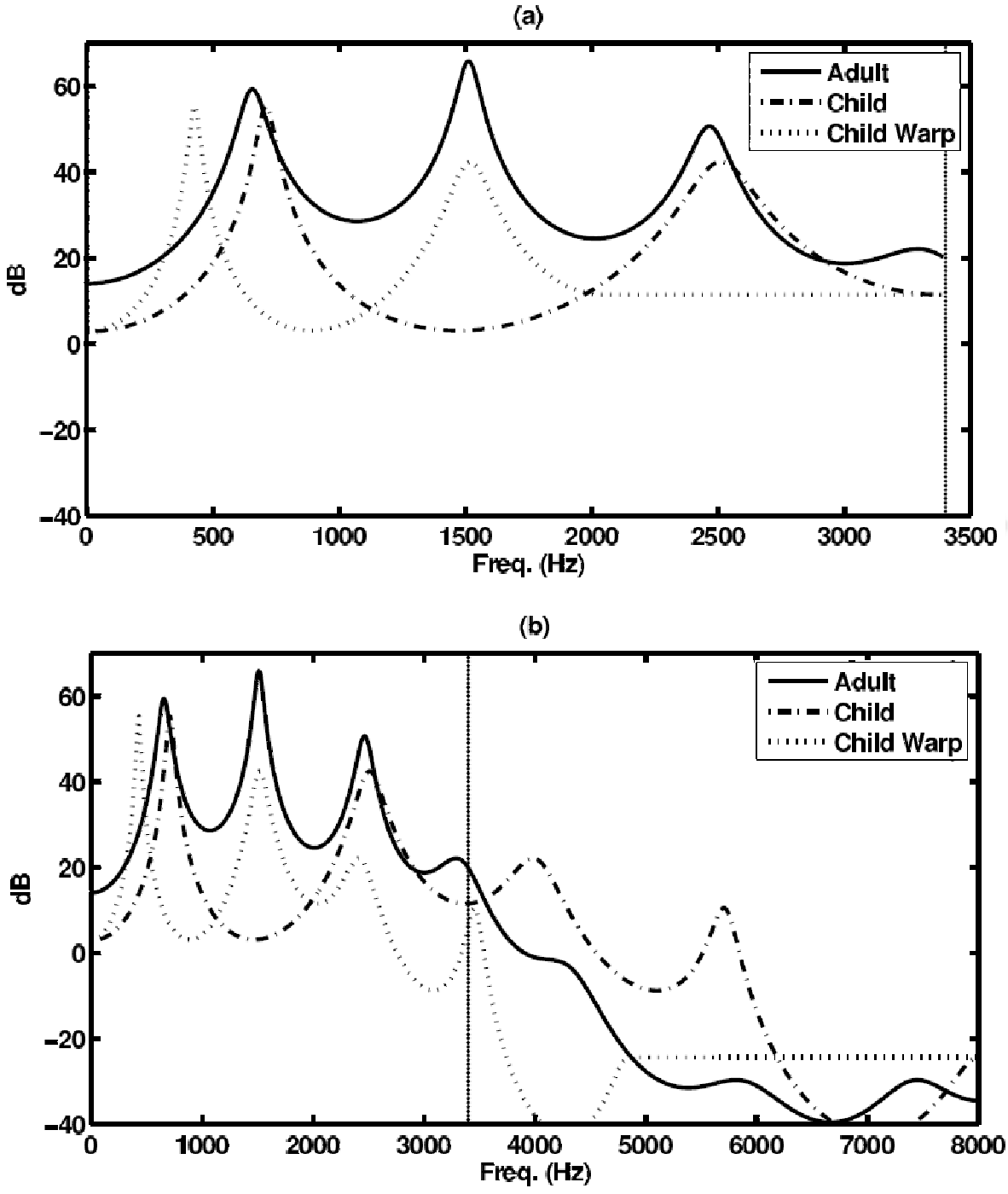

Figure 2.5: Linear predictive coding (LPC) spectra of a vowel /aa/ for an adult and a child speakers along with the warped spectrum of that child for (a) narrowband speech case, (b) wideband speech case. Note that the loss of significant higher band spectral information in the narrowband case leads to poor match between warped child's and adult's spectra unlike the wideband case.

The goal of the present work is set as follows: Given an ASR system trained using wideband adults' speech and needs to be tested using narrowband children's speech. In such a case, the performance of the ASR system degrades severally due to missing information in the higher band from 3.4-8 kHz. To minimize the mismatch, feature level ABWE methods needs to be developed for reconstructing the missing information in the range from 3.4-8 kHz. The efficacy of the developed/proposed methods will be demonstrated by conducting ASR studies.

To carry out the proposed goal, the present work is organized as follows: The first step is to develop an ASR system using standard database, features and methodology which can be used for comparing various ABWE methods that will be developed. TI-DIGITS speech corpus is used for the present work. A digit recognition ASR system is developed using MFCC features and HMM modelling. The different systems developed include narrowband adults' speech trained and wideband adults' speech trained ASR systems. These are tested using different types of testing speech to experimentally observe the degradation in the performance when children's testing speech is used. The next issue to experimentally study is the significance of ABWE. For this an existing standard method for ABWE is implemented and narrowband children's speech is subjected to ABWE. The ABWE children's speech is then used for testing the adults' trained systems. If the ASR system shows significant improvement in performance, then it supports our choice of direction for the work.

There are several attempts in the literature for reducing the mismatch between children's speech and adults' speech interacting technologies. The prominent attempts include those based on VTLN. A standard VTLN approach is implemented and both adults' and children's speech are subjected to VTLN process. The VTLN processed speech is tested with adults' trained models. The performance comparison between ABWE and VTLN cases helps in understanding the effectiveness of both the approaches. Also, a combination of the two may give an idea about how both can help in improving the performance of ASR system under mismatched condition. Another attempt is handling pitch mismatch by truncation [54]. The same can be implemented and truncated coefficients cases can be compared with ABWE case to understand the differences in both the cases. All these studies will establish the ground work for the proposed goal of

ABWE for children's speech recognition under mismatched condition. Therefore the above mentioned issues are described in Chapter 3.

The Gaussian mixture model (GMM) can be used for the modelling of joint probability density function (PDF). The joint modelled GMM can then be used for obtaining information with respect to any one PDF. This property can be exploited for ABWE. During training the GMM, the joint PDF of narrowband and corresponding wideband features can be constructed. Then during testing, given narrowband component features, the wideband features can be estimated using the joint PDF. An ABWE method based on this thought process is described in Chapter 3. The process can be carried out in a global or class-specific sense and both are developed. The GMM based on class-specific approach may give improved performance. The analysis of performance for different cases is done using the information-theoritic approach based on mutual information.

The previous work demonstrated how a joint PDF modelling technique like GMM can be used for extending the bandwidth of narrowband children's speech suitable for ASR. The motivation is more from the capability of GMM approach for the creation of joint PDF and less from speech-specific knowledge. The joint PDF created during training is used during testing phase of narrowband children's speech to obtain corresponding wideband speech. A bandwidth extension can also be developed by exploiting speech-specific knowledge at the feature level. In the existing ABWE extension techniques, the focus is mainly on synthesizing wideband speech from narrowband speech using source-filter model. Invariably, this approach involves obtaining information about LP coefficients for the high frequency range and using for synthesis. Alternatively, the current focus is mainly on ASR. Therefore, as long as the derived feature improves information in the higher frequency range that is useful for ASR, it will suffice. Along these lines, MFCC is the mostly used feature for ASR. Therefore is it possible to develop an ABWE method suitable for ASR using MFCC? The existing work of ABWE using MFCC focuses on source-filter framework. The proposed method only involves generating MFCC representing high frequency range information and using them along with narrowband MFCC for ASR. A method can be developed for ABWE using MFCC as follows:

During training, the wideband children's speech is taken, MFCC is extracted separately for low frequency and high frequency ranges and a joint PDF is developed using these two set of MFCCs. During testing, from the narrowband children's speech, the low frequency range MFCC are extracted and applied to the joint PDF to derive the corresponding high frequency range MFCC. Both these MFCCs are concatenated and used as features for ASR. Accordingly, improvement in ASR may be expected. Also, the significance of other information like age, dynamics associated with vocal tract characterized by the delta features and VTLN can be used along with MFCC for further refinement of ASR performance.

The sparse representation based signal modelling approach is found to result in state of the art performances in many signal processing applications [130–133]. Motivated by this, the present work also explores developing ABWE methods exploiting the sparse representation modelling of speech frames in the signal domain. The main merit of this approach is it is a non-parametric ABWE approach and also does not make use of the mostly followed source-filter model approach for ABWE. The sparse representation aims in obtaining a sparse vector for the given feature vector using a dictionary. The given vector is of low dimension and the obtained sparse vector is a very high dimension vector and accordingly most of the elements in the sparse vector will be zero or negligible. Hence the name of the resulting vector as sparse vector. The basis for using sparse representation framework is as follows: The conventional Fourier based dictionaries have orthogonal bases and those have highly narrowband spectral characteristics. As a result of this, although these dictionaries happen to produce very accurate representation of speech signals, but their representations are neither sparse nor have much similarity for wideband and narrowband speech. In contrast, if we create a dictionary that produces a sparse representation for speech signals, then it is more likely to produce similar representations for wideband and narrowband speech. Motivated by this hypothesis, ABWE methods are explored in which the sparse representation of the narrowband speech obtained with respect to a narrowband dictionary are applied to a corresponding wideband dictionary to achieve the reconstruction of high band information for the given narrowband signal. The KSVD algorithm is used for a creating a learned redundant dictionary for sparse representa-

tion [132]. The wideband dictionaries created separately for voiced and unvoiced speech signals are decimated and interpolated to obtain narrowband dictionaries. Using both of them the ABWE methods are developed. Several refinements to the basic approach like linear transformation of narrowband interpolated sparse coefficients, lookup constrained linear transformation and semi-coupled dictionary are explored.

3

Artificial Bandwidth Extension for Speech Recognition

Contents

The first step is to understand the significance of ABWE for speech recognition, especially, children's speech recognition under mismatched condition. This is because, the following doubts arise when ABWE framework is proposed. How effective is ABWE for speech recognition? That is, whether it will add additional information to the existing one or not. The next doubt is, to take care of mismatch in the children's speech recognition task, already methods based on VTLN and pitch mismatch compensation are proposed in the literature, then how different the proposed ABWE framework compared to these. To answer these set of doubts, a series of experiments are designed, experimentally studied and described in this chapter.

An ASR system using TI-DIGITS database as a baseline system is described first. The implementation of existing ABWE method and using it for ASR study is described next. The ASR study using existing VTLN approach is then described. The ASR study based on reducing pitch mismatch using cepstral truncation is explained next. While performing ASR studies using VTLN and cepstral truncation, the same are compared with ABWE to understand the similarities and differences.

In this chapter, as a first study, the development of an ASR system using standard database, features and methodology is described. Such a system can be used for comparing various ABWE methods. TI-DIGITS speech corpus is used for the present work. A digit recognition ASR system is developed using MFCC features and HMM modelling. The different systems developed include narrowband adults' speech trained and wideband adults' speech trained ASR systems. These are tested using different types of testing speech to experimentally observe the degradation in the performance when children testing speech is used.

The next study focused on experimentally studying the significance of ABWE. For this, implementation of existing method for ABWE based on linear prediction (LP) analysis in source-filter framework is described. The narrowband children's speech is then subjected to ABWE using the implemented ABWE method. The ABWE children's speech is then used for testing the adults' trained system.

A standard VTLN approach implemented in HTK toolbox based on linear warping is used for understanding the significance of the same for children's speech recognition under mis-

matched condition. Both the narrowband and wideband children's and adults' speech are subjected to VTLN and the normalized speech signals are used for testing adult trained models. Also, a study that combines both ABWE and VTLN is described next.

The earlier attempt for handling pitch mismatch by truncation is described next [54]. The same is implemented and truncated coefficients cases is compared with ABWE case to understand the differences and also potentials of both directions. As a result a combination of the two is proposed to demonstrate the significance of ABWE along with existing pitch mismatch handling case. All these studies are aimed at establishing the ground work for the proposed goal of ABWE for children's speech recognition under mismatched condition.

The rest of the chapter is organized as follows: Section 3.1 describes the development digit recognition system using TI-DIGITS database by extracting MFCC features and modelling by HMM for children's and adults' speech. The implementation of standard ABWE method based on LP analysis and the corresponding ASR study using the same is described in Section 3.2. Section 3.3 describes studies related to VTLN for children's speech recognition under mismatched condition and its comparison to earlier ABWE based ASR study. The implementation of cepstral truncation based method for handling pitch mismatches and the corresponding ASR study along with comparison with ABWE are described in Section 3.4. The summary and conclusions of different studies performed in this chapter and the scope of the next chapter are described in Section 3.6.

3.1 Development of baseline speech recognition system

This section describes the process of development of baseline ASR system that will be used for studying the children's speech recognition under mismatched condition. For the present work, a standard database that contains recordings from both children and adult speakers under similar conditions is needed. Accordingly, even though there many databases for the ASR study, TI-DIGITS database best suits our requirement. Accordingly, digit recognition task is used as ASR study in the present work.

3.1.1 TIDIGITS corpus

The TI-DIGITS corpus was produced by Texas instruments in 1984 and the salient features about the same are as follows [134]. It is a large speech database that can be used for speaker independent digit recognition task. It is a dialect balanced database consisting of more than 25 thousand digit sequences spoken by 326 men, women and children. The speech data was collected in a quiet environment and digitized at 20 kHz. Out of the 326 speakers, there are 111 men, 114 women, 50 boys in the age group 6-14 years and 51 girls in the age group 8-15 years. The utterances collected are digit sequences. Eleven digits were used: zero, one, ..., nine and oh. Seventy-seven sequences of these digits including 22 isolated, 11 two-digit, 11 three-digit, 11 four-digit, 11 five digit and 11 seven digit sequences were collected for each speaker. Hence each speaker provided 253 digits and 176 digit transitions.

For the present work, the whole database originally sampled at 20 kHz was resampled to 16 kHz sampling frequency for obtaining wideband speech database and also was resampled to 8 kHz sampling frequency for obtaining narrowband speech database. The adults' and children's speech database is further subdivided into training and testing sets. The adults' training set includes a total of 197 speakers having 35,566 digits and totalling to about 5.3 hrs of speech data. The adults' testing set includes a total of 81 speakers having 10,813 digits resulting in about 1.6 hrs of test speech. The children's speech consists of 64 speakers for training and 49 speakers for testing. The children's training database consists of 14,725 digits leading to about 4.4 hrs of training speech and the children's testing database consists of 10,800 digits leading to about 1.9 hrs of testing speech.

3.1.2 Design of ASR studies

The TI-DIGITS database as described above composes of both adults' and children's speech. For performing the current work of children's speech recognition under mismatched condition, unless specified, mismatched condition refers to the case of acoustic mismatch. This particularly happens when adults' and children's speech are considered. The following ASR studies are designed to understand the effect of mismatch in terms of narrowband and wideband speech,

and also adults' and children's speech. The first set of studies include *matched condition* where the models are trained using adults' speech and also testing using adults' speech. In the matched condition itself, we can have either narrowband or wideband cases. The second set of studies include *mismatched condition* where the models are trained using adults' speech and tested using children's speech. Further, in the mismatched condition, we can have either narrowband or wideband cases. Finally, the term children's speech recognition under mismatched condition refers to the case of training using wideband adults' speech and testing using narrowband children's speech.

3.1.3 Feature extraction

The speech is pre-emphasized using a factor of 0.97. For short term processing, speech is processed in frames of 25 ms with a frame rate of 100 Hz. For each frame, a 21 channel filterbank (between 0.3-3.4 kHz) is used for narrowband signals while 26 channel filterbank (between 0-8 kHz) is used for wideband signals. The 13 dimensional MFCC (C_0-C_{12}) forms the base feature vector in both cases. In addition to the base features, their first and second order derivatives are also appended making the final dimension as 39. Cepstral mean subtraction is also applied to all features.

3.1.4 Digits models

The connected digit recognition task used in this work has been developed using the HTK toolkit [72]. The 11 digits (0-9 and OH) are modelled as whole word left-to-right hidden Markov model (HMM). Each word model has 16 states with simple left-to-right paths and no skip paths over the states. The observation densities are mixture of five multivariate Gaussian distributions with diagonal covariance matrices. The silence is explicitly modelled using three state HMM model having six Gaussian mixtures per state. A single-state short pause model tied to the middle state of the silence model is also used.

3.1.5 Performance Evaluation

The recognition performance is evaluated using different testing sets and word error rate (WER in %) is used as a metric for comparison of different systems. The performance of different digit recognition systems are tabulated in Table 3.1, where the trained models are using adults' speech and testing using either adults' or children's speech. Further the testing case can include either narrowband or wideband cases. The narrowband adults' trained system gives a WER of 0.44 % for narrowband adults' test speech. This result shows that the training and testing cases are acoustically matched well. Alternatively, testing with narrowband children's speech results in a WER of 9.37 %. The very poor WER infers the significant acoustic mismatch present between adults' and children's speech. The wideband adults' trained system gives a WER of 0.35 % which is a marginal improvement over narrowband adults' testing case. Alternatively, testing with wideband children's speech results in a WER of 3.21 %. The WER is a significant improvement over narrowband children's testing case indicating the significant information in the children's speech is present in the higher frequency range (3.4 to 8 kHz). These set of experiments demonstrate the need for reconstructing the information in the higher frequency range of children's speech.

Table 3.1: *Recognition performances for adults' (AD) speech and children's (CH) speech test sets having narrowband (NB) and wideband (WB) speech data of the TI-DIGITS corpus.*

Test set	WER%	
	NB	WB
	Base	Base
AD	0.44	0.35
CH	9.37	3.21

3.2 Proposed Spectral Loss Compensation for ASR using ABWE

The earlier study demonstrated the need for the information present in the high frequency range (3.4-8 kHz) to obtain significant performance improvement in case of children's speech recognition under mismatched condition. To see how much we really benefit if we have an

ABWE extension method, this section describes the study using an existing ABWE method.

Most of the ABWE algorithms for speech signal are based on the source-filter model of the speech production. The estimation of the missing higher band signal components is performed in a two-stage procedure, indirectly via the model of source-filter: in the first step the wideband source signal is estimated from the narrowband source signal. In the second step higher band filter parameters are estimated from narrowband filter parameters and then used in combination with narrowband filter parameters to determine the wideband filter parameters. This approach is in general well suited for the extension of both high frequencies and low frequencies for the given 0.3-3.4 kHz signal.

In this work for the extension of bandwidth of narrowband signal, we have used an algorithm proposed in [12]. In the following, we describe the basics of GMM based ABWE algorithm used. The generic block diagram of ABWE algorithm is shown in Figure 3.1. According to the structure of source-filter model, the bandwidth extension is performed separately for the excitation source signal and the filter representing the spectral envelope of the speech signal. Since these two constituents of the speech signal can be assumed to be mutually independent to a certain extent, separate optimization of these two parts of the algorithm leads to an approximation of the global optimum. The nonlinear operation used ... The LP analysis is performed after nonlinear to derive fullband LP residual.

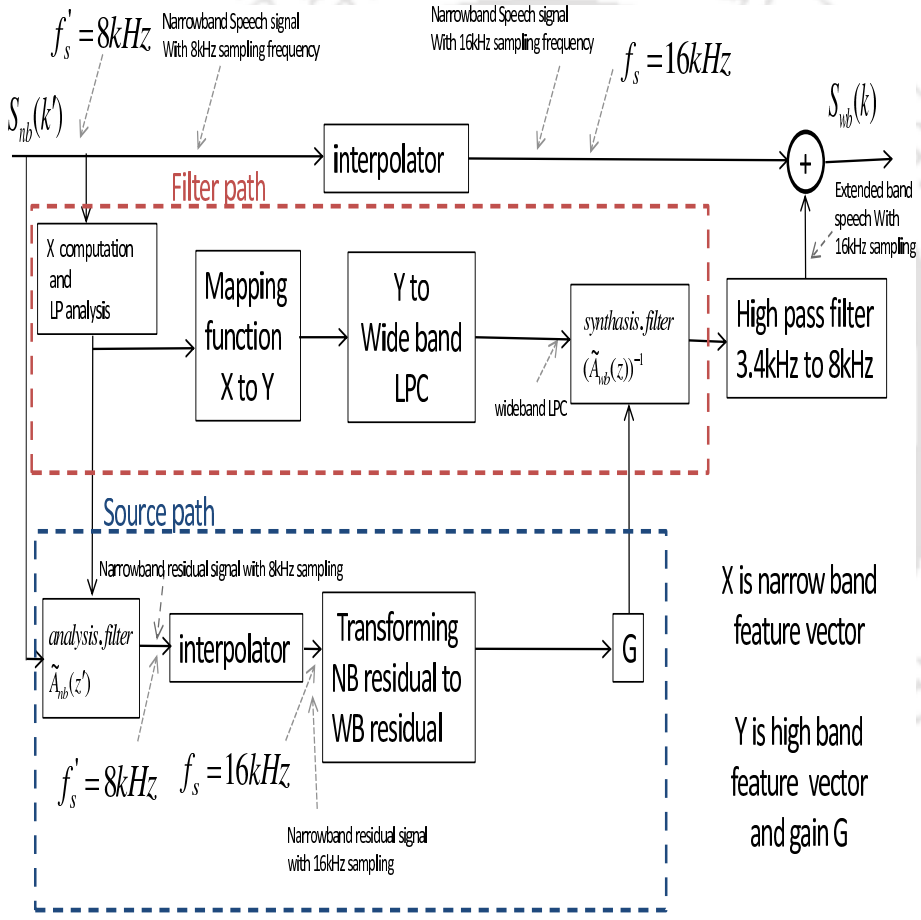

Figure 3.1: *Block diagram of the source-filter based generic ABWE algorithm.*

As shown in Figure 3.1, $\mathbf{X}$ represents the feature for the narrowband signal. As described in [12], the feature $\mathbf{X}$ consists of two broad components $\mathbf{X_{scr}}$ and $\mathbf{X_{acf}}$. $\mathbf{X_{scr}}$ is having five scalar features and $\mathbf{X_{acf}}$ having ten energy normalized auto-correlation coefficients as defined in Eq (3.1), (3.2) and (3.3).

$$\mathbf{X} = [\mathbf{X_{acf}}^T, \mathbf{X_{scr}}^T]^T \tag{3.1}$$

$$\mathbf{X_{scr}} = [X_{zcr}, X_{gi}, X_{nrf}, X_k, X_{sc}]^T \tag{3.2}$$

$$\mathbf{X_{acf}} = [X_{acf}(1), X_{acf}(2), X_{acf}(3), ..., X_{acf}(10)]^T \tag{3.3}$$

where X_{zcr}, X_{gi}, X_{nrf}, X_k and X_{sc} denote the zero crossing rate, gradient index, normalized relative frame energy, local kurtosis and spectral centroid, respectively.

$\mathbf{Y}$ represents the feature for the higher band portion of the wideband speech signal. It consists of the logarithm of the ratio of narrowband (nb) and higher band (hb) energies and the rest ten coefficients are the selective linear prediction cepstral coefficients (SLPCC) computed from the higher band (3.4-8.0 kHz) spectral information of the signal as described in [100].

$$\mathbf{Y} = [G, C_1, C_2, ..., C_{10}]^T, \text{where } G = \log(\frac{\sigma_{nb}}{\sigma_{hb}}) \tag{3.4}$$

It should be emphasized at this point that both X and Y are composite features having different information as described in the above equations. Modulation technique is used to transform NB residual to WB residual signal.

3.2.1 Selective linear prediction

Selective linear prediction (SLP) is the linear prediction process performed on a limited frequency band of a signal. In our case, it is used to model the extended band from the lower limit Ω_l or k_l to the upper limit Ω_u or k_u. One method of performing linear prediction on a limited frequency interval could be to band pass filter the input signal and then do conventional LP analysis. The band pass filter should have the above mentioned frequencies as cut-off frequencies. This method has one major weakness, namely, the linear prediction will try to model the slopes of the band pass filter, which result in wasting model capabilities. To avoid

this, SLP is proposed. The SLP algorithm involves the following four steps [135, pp. 148]:

Compute the power spectrum: compute the Fourier transform of the signal and square its magnitude spectrum:

$$P(k) = |DFT\{s(n)\}|^2 \tag{3.5}$$

where, DFT is N point discrete Fourier transform.

Create a translated spectrum: Form a translated power spectrum $P_t(m)$ as shown in Figure 3.2, by letting $P_t(m) = P(k)$ where $k = [k_l, k_{l+1}, \ldots, k_u, N - k_u, \ldots, N - k_l]$ and $m = [0, ..., 2(k_u - k_l) - 1]$, An example of the translated power spectrum is shown in Figure 3.2.

Compute the autocorrelation sequence: Do an inverse DFT using $N_1 = 2(k_u - k_l)$ point DFT to obtain the autocorrelation function of the translated spectrum.

$$r_t(n) = IDFT\{P_t(m)\} \tag{3.6}$$

Compute the SLP model parameters: Based on the autocorrelation function, calculate the LP coefficients using the Levinson Durbin algorithm [136, pp. 278-280].

The envelope can be represented by the LP coefficients, which describe the all-zero filter $A(z)$ (the analysis filter) defined as [135]:

$$A(z) = \sum_{i=0}^{p} a_i z^{-i} \tag{3.7}$$

The filter coefficient $(a_0, ..., a_p)$ defines the p^{th} order analysis filter. Linear prediction is done by minimizing the prediction error between the actual sample $s(n)$ and the predicted sample $\hat{s}(n)$, based on an optimization criterion.

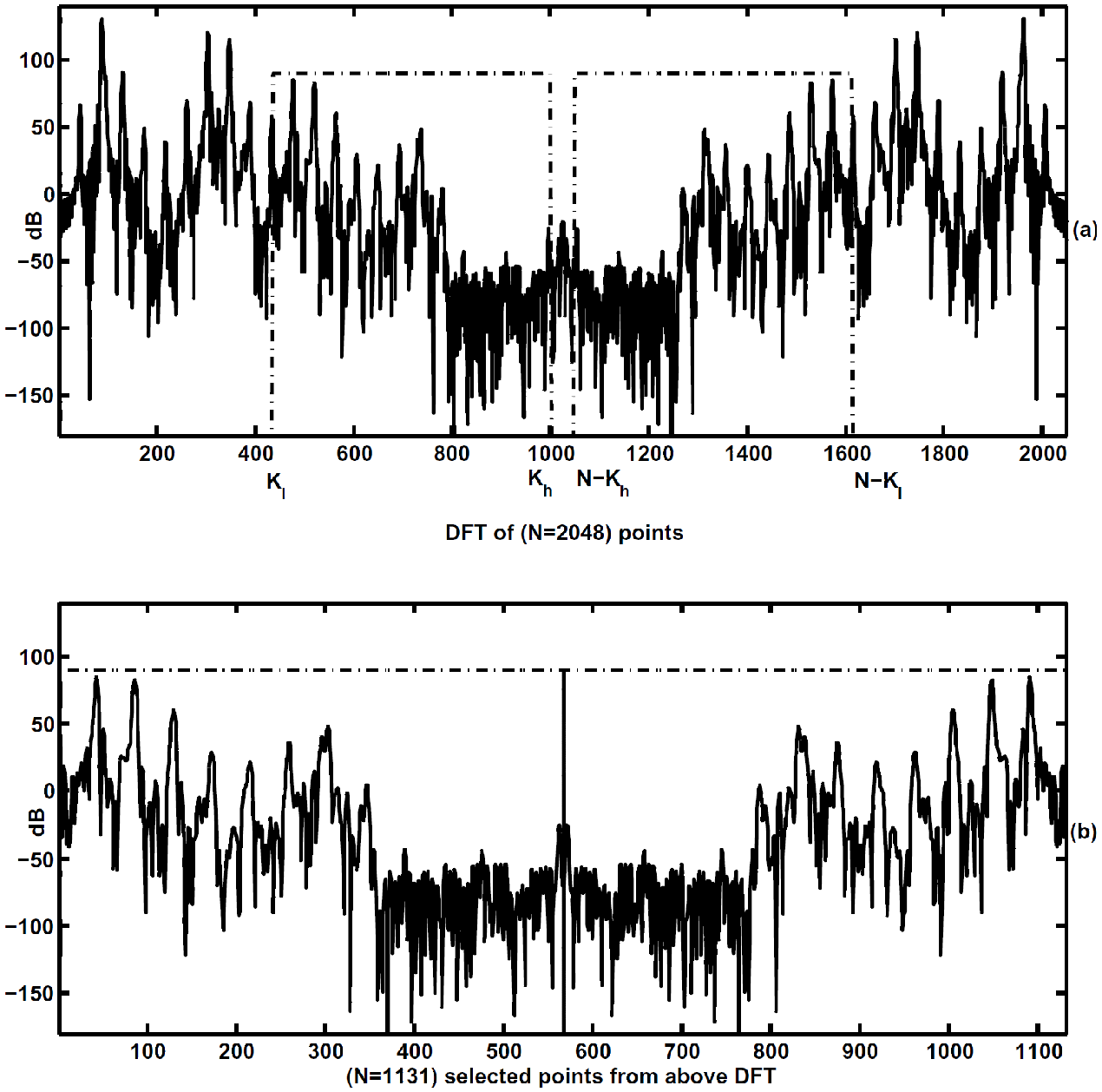

Figure 3.2: *The wideband power spectrum ($P(k)$) in (a) and the translated spectrum ($P_t(m)$) in (b). It should be noted that the spectra are double-sided. Portion of spectrum selected using dotted line in (a) is translated to complete range of normalized frequency in (b)*

The prediction error is defined as [135]:

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{i=1}^{P} a_i s(n-i) \tag{3.8}$$

Usually the total squared error (α) is minimized, which is:

$$\alpha = \sum e(n)^2 \tag{3.9}$$

This can be done by using e.g. the autocorrelation method [135, 137].

The LP coefficients are a simple representation of the envelope and are computationally attractive. The LP coefficients are highly correlated among each other. If a single LP coefficient are altered, it has high impact on the shape of complete envelope. Two sets of LP coefficients with only one coefficient being different have a small Euclidean distance. The shape of the envelopes however may differ a lot. Quantifying the LP coefficients from an Euclidean distance would therefore not be appropriate if similar envelopes should be clustered together.

Cepstral coefficients from LP coefficients: The common method to calculate the real cepstral coefficients is to perform a discrete time Fourier transform of a signal, apply the log operator to the magnitude of output transformation, and then do an inverse discrete time Fourier transform.

In this thesis, we take advantage of the possibility of deriving the cepstral coefficients from the LP coefficients. These cepstral coefficients also popularly called as LP cepstral coefficients (LPCC). Markel and Gray [135, pp. 230] have an equation from which both the cepstral coefficients $c(n)$ and the LP coefficients a_i can be derived recursively:

$$-n\ c(n) - n\ a_n = \sum_{k=1}^{n-1} (n-k)\ c(n-k)\ a_k \text{ for } n > 0 \tag{3.10}$$

where,

$$a_0 = 1 \text{ and } a_k = 0 \text{ for } k \geq p \text{ and } c(0) = \ln(\alpha) \tag{3.11}$$

p is the order of the autoregressive model and α is the gain coefficient. In order to find the cepstral coefficients Eq 3.10 is rewritten to:

$$c(n) = \frac{-1}{n}(n\ a_n + \sum_{k=1}^{n-1}(n-k)\ c(n-k)\ a_k) = -a_n - \sum_{k=1}^{n-1}\frac{(n-k)}{n}\ c(n-k)\ a_k, \text{ where } n > 0 \tag{3.12}$$

The cepstral coefficients $c(0)$ and $c(1)$ can be found from 3.11 and 3.12 as:

$$c(0) = \ln(\alpha)\ c(1) = -a_1 \tag{3.13}$$

The derived cepstral coefficients can be converted back into LP coefficients and again by applying the following ($n > 0$):

$$a_n = \frac{-1}{n}\ (n\ c(n) + \sum_{k=1}^{n-1}(n-k)\ c(n-k)\ a_k) = -c(n) - \sum_{k=1}^{n-1}\frac{(n-k)}{n}\ c(n-k)\ a_k \tag{3.14}$$

The conversion between LP coefficients and CC is relatively cheap and easy to perform. The conversion has the advantage that it can be reversed without any loss in information. It is an advantage to transform the LP coefficients into cepstral coefficients, because the cepstral representation has the advantage of having sufficient decorrelation among the coefficients [2, pp. 212]. This is advantageous for modelling cepstral coefficients with a probability density function (PDF), because the parameters of the PDF, can be estimated (optimized) for one element in the feature independently of the other elements. Furthermore minimizing the Euclidean distance between two sets of cepstral coefficients, is closely related to minimizing the log spectral distortion (LSD) between the two spectra [135].

3.2.2 Gaussian mixture model

Let $\mathbf{X} \in \mathbb{R}^k$ be the feature vector of narrowband speech and $\mathbf{Y} \in \mathbb{R}^l$ be the feature vector of the higher band speech. Then vector $\mathbf{Z} = \left[\mathbf{X^T}, \mathbf{Y^T}\right]^T \in \mathbb{R}^n$ is formed and is modelled using the Gaussian mixture model (GMM) to estimate the joint PDF of $\mathbf{X}$ and $\mathbf{Y}$ as depicted in Figure 3.3.

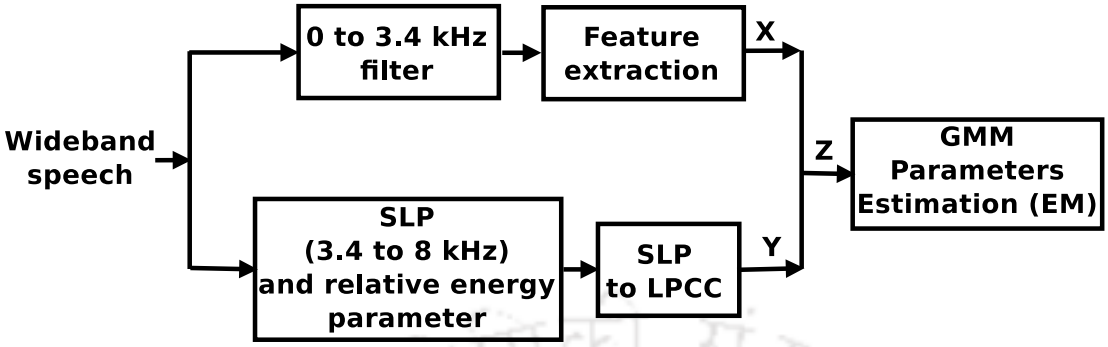

Figure 3.3: *Block diagram illustrating the generation of features of the narrowband and the higher band portions of the signal and modelling of their joint PDF using GMM.*

The PDF of $\mathbf{Z}$ is modelled as a mixture of M n-variate Gaussian PDFs,

$$P(\mathbf{Z}|\lambda) = \sum_{i=1}^{M} \alpha_i b_i(\mathbf{Z}), \quad \sum_{i=1}^{M} \alpha_i = 1 \text{ and } \alpha_i \geq 0 \tag{3.15}$$

where $b_i(\mathbf{Z})$ and $\alpha_i, i = 1, ..., M$ are the component densities and the component weights, respectively. Each component density is a n-variate Gaussian function of the form:

$$b_i(\mathbf{Z}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{C}_{zz_i}|^{\frac{1}{2}}} e^{-\frac{1}{2}\left[(\mathbf{Z}-\boldsymbol{\mu}_{z_i})^T \mathbf{C}_{zz_i}^{-1} (\mathbf{Z}-\boldsymbol{\mu}_{z_i})\right]} \tag{3.16}$$

with $\boldsymbol{\mu}_{z_i}$ is $n \times 1$ mean vector and $\mathbf{C}_{zz_i}$ is $n \times n$ covariance matrix with

$$\boldsymbol{\mu}_{z_i} = \begin{bmatrix} \boldsymbol{\mu}_{x_i} \\ \boldsymbol{\mu}_{y_i} \end{bmatrix}$$

and

$$\mathbf{C}_{zz_i} = \begin{bmatrix} \mathbf{C}_{xx_i} & \mathbf{C}_{xy_i} \\ \mathbf{C}_{yx_i} & \mathbf{C}_{yy_i} \end{bmatrix}$$

The parameters of GMM model, $\lambda = \{\alpha_i, \boldsymbol{\mu}_i, \mathbf{C}_i\}$, are estimated using the expectation and maximization algorithm.

3.2.3 Spectral envelope reconstruction

Given the joint PDF of $\mathbf{X}$ and $\mathbf{Y}$, the goal is to find a mapping function $F(.)$ that minimizes the mean square error,

$$\varepsilon_{\text{mse}} = E\left[|\mathbf{Y} - F(\mathbf{X})|^2\right] \tag{3.17}$$

where $E[.]$ denotes expectation, and $F(\mathbf{X})$ is the reconstructed higher band feature vector. From the Bayesian estimation theory, the mapping function minimizing the mean squared error between the reconstructed higher band features and the original higher band features turns out to be the mean of the conditional PDF $P(\mathbf{Y}|\mathbf{X})$ and is given by the regression,

$$F(\mathbf{X}) = E[\mathbf{Y}|\mathbf{X}] = \sum_{i=1}^{M} h_i(\mathbf{X}) \left[\boldsymbol{\mu}_{y_i} + \mathbf{C}_{yx_i}\mathbf{C}_{xx_i}^{-1}\left(\mathbf{X} - \boldsymbol{\mu}_{x_i}\right)\right] \tag{3.18}$$

where

$$h_i(\mathbf{X}) = \frac{\frac{\alpha_i}{(2\pi)^{\frac{n}{2}}|\mathbf{C}_{xx_i}|^{\frac{1}{2}}} e^{-\frac{1}{2}\left[\left(\mathbf{X}-\boldsymbol{\mu}_{x_i}\right)^T \mathbf{C}_{xx_i}^{-1}\left(\mathbf{X}-\boldsymbol{\mu}_{x_i}\right)\right]}}{\sum_{j=1}^{M} \frac{\alpha_j}{(2\pi)^{\frac{n}{2}}|\mathbf{C}_{xx_j}|^{\frac{1}{2}}} e^{-\frac{1}{2}\left[\left(\mathbf{X}-\boldsymbol{\mu}_{x_j}\right)^T \mathbf{C}_{xx_j}^{-1}\left(\mathbf{X}-\boldsymbol{\mu}_{x_j}\right)\right]}} \tag{3.19}$$

the weighting function $h_i(\mathbf{X})$ denotes *a posteriori* probability that ith Gaussian component has generated the narrowband feature $\mathbf{X}$.

Finally, the estimated higher band feature vector is converted to corresponding spectral magnitude spectra through cepstral to LPC recursion before appending it to the original narrowband magnitude spectrum for bandwidth extension purpose.

3.2.4 ASR Study using ABWE

For the ABWE algorithm described above, the different parameters chosen are: window length of 20 ms, window shift of 10 ms, the narrowband LPC order of 10, the higher band LPC order of 20 and the GMM of size 512. For extension of the bandwidth of adults and children's narrowband speech test sets, separate GMM models were trained from appropriate condition training data, mutually exclusive to the test sets. The recognition performance of the bandwidth extended narrowband test sets are evaluated on the connected digit recognition system described earlier, developed using the wideband speech data from TI-DIGITS. The performance of ASR study for ABWE case is summarized in Table 3.2. The ABWE adults' speech gives a WER of 0.57 % which is slightly inferior compared to 0.35 %. Also for the NB adults' case itself the WER is 0.44 %. This comparison infers that, in case of adults' speech,

Table 3.2: *Recognition performances for adults' (AD) speech and children's (CH) speech test sets having narrowband (NB) and wideband (WB) speech data of the TI-DIGITS corpus.*

Test set	WER%		
	NB	WB	ABWE
	Base	Base	Base
AD	0.44	0.35	0.57
CH	9.37	3.21	4.06

most of the information is the narrowband itself. Hence the gain is minimal when we move from narrowband to wideband case. The slightly inferior performance for ABWE case can be attributed to the signal processing distortion occurring in different steps of ABWE.

The performance of ABWE children's speech tested against wideband adults' models give a WER of 4.06% which is a significant improvement over the narrowband result of 9.37%. This result infers that the acoustic mismatch between the children's and adults' speech is high and is mostly in the higher band. That is why, the ABWE is helping in improving the performance of children's speech recognition under mismatched condition. Note that, in this case also, the same signal processing distortion is introduced into ABWE children's speech. The signal processing steps essentially represent narrowband to wideband conversion processes and involve some distortion. Hence the WER is 4.06 % which could have been 3.21 % for the best case. The various results from this experiment indicate that ABWE indeed significantly helps for improving the ASR performance under mismatched condition for the children's speech.

3.3 ASR using VTLN

When there is an acoustic mismatch due to vocal tract, the most extensively used approach to minimize this mismatch is vocal tract length normalization (VTLN). The previous section demonstrated the significance of ABWE for ASR under acoustic mismatch related to vocal tract itself. Therefore, in the presence of VTLN, the need for ABWE is questioned. For this an ASR study needs to be performed using VTLN and then compare the two cases.

VTLN is a speaker normalization method in which the inter-speaker acoustic variability due to varying vocal tract lengths among speakers is reduced by warping the frequency axis of the

speech spectrum of each speaker [70, 71]. In this work, VTLN is performed on an utterance-by-utterance basis on the test speech data as described in [73].

For warping the frequency axis of the utterances during computation of MFCC features, the piece-wise linear frequency warping of filterbank, as supported in the HTK [72], is used. The spacing and the width of the filters in the mel filterbank are changed while maintaining the speech spectrum unchanged. As the warping would lead to some filters being placed outside the analysis frequency range, to avoid the same a piece-wise linear warping function of the frequency axis of the mel filterbank is employed [73]:

$$g_\alpha(f) = \begin{cases} \frac{1}{\alpha} f & 0 \leq f \leq f_c \\ \frac{1}{\alpha} f_c + \frac{f_{\max} - \frac{1}{\alpha} f_c}{f_{\max} - f_c}(f - f_c) & f_c < f \leq f_{max} \end{cases} \tag{3.20}$$

where, $f_{\max}$ denotes the maximum signal bandwidth (4 kHz in narrowband and 8 kHz in wideband) and f_c is an empirically chosen frequency of 3.4 kHz.

The optimal frequency warp factor for the test signal is estimated based on a maximum likelihood (ML) grid search over a possible range of warp factors given a current set of acoustic models under the constraint of the first-pass transcription of the test signal. In this work, for doing ML grid search, each speech feature is warped by 13 different factors ranging from 0.88-1.12 in steps of 0.02. Given the various warped features, the optimal value $\hat{\alpha}$, by which the frequency axis of speech spectrum is warped, is estimated as:

$$\hat{\alpha} = \arg\max_{\alpha} P(\mathbf{x}_i^\alpha | \lambda, W_i) \tag{3.21}$$

where, $\mathbf{x}_i^\alpha$ represents the warped feature for the ith utterance with frequency axis of speech spectrum scaled by factor α. λ represents the HMM based speech recognition model and W_i is the transcription of the ith utterance. W_i is determined by the first recognition pass using the unwarped feature set. Ideally, the effect of using an optimal scaling factor selected in this way for each utterance is that of normalizing the test speech data with respect to the average vocal tract length of the training population of the recognition model set λ, thus reducing the inter-speaker acoustic variability between the training and the test data.

Table 3.3: *Recognition performances for adults speech (AD) and children's speech (CH) test sets having narrowband (NB), wideband (WB) and artificial bandwidth extended (ABWE) speech data. For assessing the quality of the reconstructed higher band spectra in the bandwidth extended signals, the performance with applying VTLN are also given.*

Test set	WER%					
	NB		WB		ABWE	
	Base	+VTLN	Base	+VTLN	Base	+VTLN
AD	0.44	0.43	0.35	0.35	0.57	0.57
CH	9.37	1.64	3.21	0.77	4.06	1.17

The performance of ASR system for different studies related to VTLN are summarized in Table 3.5. The performance of ASR system for VTLN normalized narrowband adults' test speech against narrowband adults' models give a WER of 0.43% as against 0.44% for without normalization case. Similarly, for VTLN normalized wideband adults' test speech against wideband adults' models give a WER of 0.35%, same for without normalization case also. Thus in case of adults' speech, VTLN seems to have minimum impact. In case of VTLN normalized narrowband children's speech tested against narrowband adults' models gives a WER of 1.64% as against 9.37% for without normalization case. This result infers that the acoustic mismatch between the children's and adults' speech is high and can be significantly reduced by employing VTLN. Next in the case of VTLN normalized wideband children's speech tested against wideband adults' models gives a WER of 0.77% as against 3.21% for without normalized case. Both these results infer that VTLN helps significantly in case of children's speech recognition under mismatched condition.

The results seem to be even better compared to stand alone performance of ABWE. Based on this we are tempted to conclude that VTLN can be used instead of ABWE. However, the two processes are different. In VTLN case, only normalization is performed for the same band, where as, in case of ABWE, missing information in the high frequency is reconstructed. Accordingly, we can combine both to get improved performance. The performance of ABWE and VTLN normalized narrow band children's speech tested against wideband adults' models gives a WER of 1.17%. Only using VTLN it is 1.64% and using ABWE it is 4.06%, where as the combination provides 1.17%. This combined result infers that, using both further improves

ASR performance. Thus any new ABWE method developed will always add value to improve the children's speech recognition under mismatched condition. Hence the need for developing new ABWE methods.

3.4 ASR using Truncation of MFCC Features

The previous two sections demonstrated the significance of reducing mismatch with respect to the vocal tract information and hence improving the children's speech recognition under mismatched condition. The next important cause for mismatch between children's and adults' speech is the difference in the physiological and also dynamic characteristics of the excitation source. The earlier work focused on reducing the excitation source related mismatch and hence improving the children's speech recognition performance [1]. In this work, one important study was reducing the mismatch by the truncation of MFCC coefficients [54]. It will be better to compare the proposed ABWE framework with this study also. Due to this, the earlier work for ASR using truncation of MFCC features is described first and then compared with ABWE framework.

To explore the effect of cepstral truncation based spectral smoothing, the base features for the test data are truncated from 16 down to 2 in step of 1. As the dimensionality of the truncated test features do not match with that of the features used for training the acoustic models, the default decoding algorithm (HVite) in HTK cannot be employed for decoding purposes. To overcome this problem either the decoding algorithm has to be suitably modified or the acoustic models need to be retrained for each truncated dimension of the test features. In this work, for ease of experimentation, we retrain the models with features of same dimensionality as those of the truncated test features. The recognition performances for both the adults' and the children's test sets for different dimensions of the truncated test features on acoustic models trained with matching feature dimensions are given in Table 3.4. Although the truncation is shown in the base feature only but the truncated test features also include their first and second order derivatives.

From Table 3.4, it is noted that the recognition performance for the children's test set

Table 3.4: *Recognition performances for adults speech (AD) and children's speech (CH) test sets for varying truncated length of base MFCC feature for narrowband (NB) and wideband (WB) speech data. For recognition purpose the first and second derivatives were also appended to corresponding base features.*

Base feature length	WER%					
	NB		WB		ABWE	
	AD	CH	AD	CH	AD	CH
16	0.54	10.81	0.27	4.77	0.45	5.73
15	0.51	10.31	**0.27**	**4.08**	**0.49**	**5.25**
14	0.48	9.80	0.29	3.62	0.53	4.60
13	**0.48**	**9.72**	0.44	3.19	0.61	4.01
12	0.48	9.18	0.40	2.97	0.68	3.74
11	0.51	9.06	0.42	2.73	0.69	3.59
10	0.57	8.56	0.43	2.45	0.68	3.32
9	0.62	8.11	0.43	2.42	0.70	3.30
8	0.64	6.66	0.48	2.10	0.74	**2.92**
7	0.63	5.63	0.55	**1.98**	0.74	3.15
6	0.68	5.37	0.59	2.47	0.82	3.59
5	0.73	5.26	0.76	1.99	1.11	3.47
4	0.92	**4.47**	1.25	2.73	1.69	4.63
3	1.32	5.21	3.07	5.48	4.33	7.76
2	5.66	15.19	6.05	10.57	8.05	15.01

improves consistently with the feature truncation. On the other hand, the recognition performance for the adult's test set degrades with increasing truncation of the features. It is obvious that the degradation in the performance for the adults' test set is due to over smoothing of the spectra with increasing cepstral truncation. However, for the children test set the improvements in the performance with increase in cepstral truncation are attributed not only to the reduction in the pitch-dependent distortions in the spectral envelope but also to the reduction in the mismatch on account of the vocal tract length (VTL) differences between the adults and the children speech due to the implicit spectral smoothing. Further, to illustrate the effect of varied truncation of MFCC feature on their corresponding spectra, the plots of the smooth spectra corresponding to various truncated feature lengths for vowel /iy/ of pitch 300 Hz are shown in Figure 3.4. It is observed that with increasing truncation, the distortions in the spectral envelope are smoothed out. But for higher truncation the spectral peaks (formants) also get smoothed out, resulting in reduced mismatch due to VTL differences between the adults and

the children's speech.

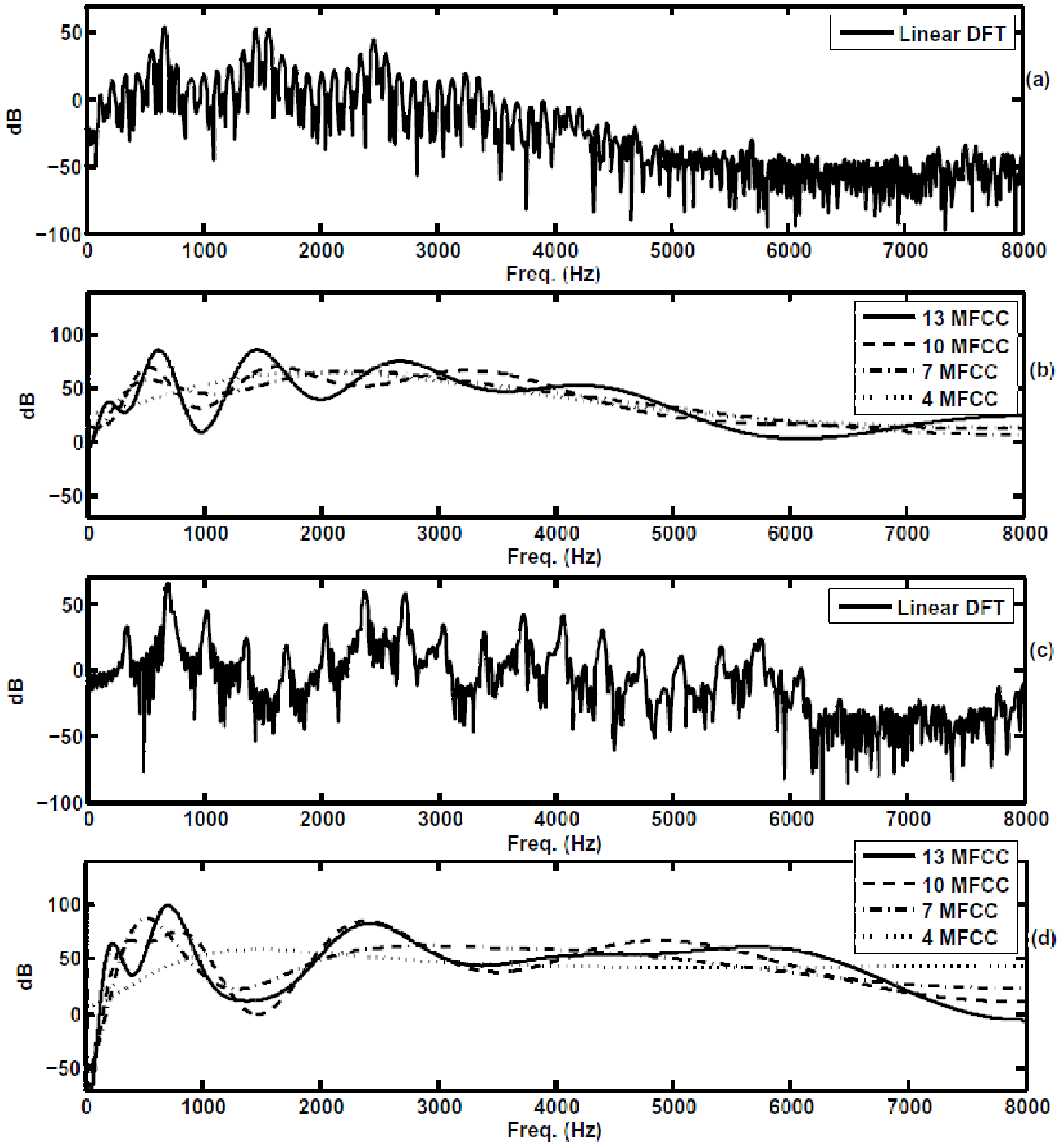

Figure 3.4: *Plots demonstrating the effect of cepstral smoothing in case of adults' and children's speech for a vowel /aa/. (a) Linear DFT spectrum for an adult speaker having pitch value of around 100 Hz (b) Smoothed mel spectra corresponding to the base MFCC features of different dimensions for that adult speaker (c) Linear DFT spectrum for a child speaker having pitch value of around 300 Hz (d) Smoothed mel spectra corresponding to the base MFCC features of different dimensions for that child speaker.*

In narrowband speech case, for 13 dimensional MFCC, the adults' test speech gives a WER of 0.48% and the children's test speech gives 9.72%. The significantly poorer performance in case of children's speech may be attributed to the mismatch due to both vocal tract as well as pitch. As the number of coefficients truncated increases, the mismatch reduces in case of children's speech. The best performance is for the 4 dimensional MFCC case where its WER is 4.47%. This shows that the children's speech recognition under mismatched condition can also be significantly improved by coefficient truncation. As described above, the coefficient truncation can be attributed to reducing mismatch with respect to both pitch as well as vocal tract length variations. In wideband speech case, the best WER for adults' case is 0.27% and the corresponding children's speech WER is 4.08% for 15 dimensional MFCC. The significant improvement in WER from 9.72% to 4.08% can be attributed to the information present in the higher frequency range of wideband speech. The performance of children's speech can be further improved in this case also by coefficient truncation. The best performance is 1.98% when only 7 dimensional MFCC is used. This study demonstrates that even in wideband case, the coefficient truncation helps. Hence we can use ABWE along with coefficient truncation.

In the next study, the earlier described ABWE method is employed to get ABWE extended speech from the narrowband speech. The ABWE extended adults' speech further decreases performance compared to narrowband case inferring the effect of signal processing distortion due to ABWE. Alternatively, the children's speech case shows significant improvement over narrowband case. The best performance is 2.92% when 8 coefficients are used. This study demonstrates that ABWE helps on top of existing coefficient truncation in case of children's speech recognition under mismatched condition.

3.5 Combining ABWE, VTLN and Cepstral Truncation Approaches

As described earlier, the basis of each of the approaches for reducing mismatch between adults' and children's speech are different. Therefore combination of them may improve the performance. The results obtained for different cases are tabulated in Table 3.5. There are

minor variations in the WER values of this study compared to the earlier sections. This is because, in the combining experiment set up, the HMM models are initially generated using MFCC of 16 dimension along with delta, and double-delta features. For performing cepstral truncation experiments, only the mean and variance vectors in each state of HMM are truncated. However, the transition matrix is kept same as in the 16 dimension case. As established earlier, there are no benefits of VTLN, ABWE or coefficients cepstral truncation in case of adults' speech data testing. However, the significance of them can be seen while using children's speech data testing. The case of children's NB speech provides the worst Performance of 9.72%. With the help of combining different approaches the best performance of 0.97% is achieved for the same children's NB speech data. The process of improving the performance by different combinations is explained below.

Table 3.5: *Recognition performances for adults' speech (AD) and children's speech (CH) test sets having narrowband (NB), wideband (WB), artificial bandwidth extended (ABWE), vocal tract length normalization (VTLN) and MFCC cepstral truncation (Trunc) speech data. Experiments are performed using the models trained for truncation experiments.*

Test set	WER%							
	NB		WB		ABWE			
	Base	Base +VTLN	Base	Base +VTLN	Base	Base +VTLN	Base +Trunc	Base +Trunc +VTLN
AD	0.48	0.47	0.44	0.40	0.61	0.55	0.58	0.50
CH	9.72	1.64	3.19	0.77	4.01	1.06	2.76	0.97

3.5.1 ABWE and VTLN

The performance of VTLN seem to be even better compared to stand alone performance of ABWE. Based on this, it may be tempting to conclude that VTLN can be used instead of ABWE. However, the two processes are different. In VTLN case, only normalization is performed for the same band, where as, in ABWE, missing information in the high frequency is reconstructed. Accordingly, we can combine both to get improved performance. The children's NB speech is subjected to ABWE and then VTLN is performed. Only using ABWE it is 4.01%, where as the combination provides 1.06%. This shows the efficacy of combining ABWE and

VTLN approaches.

3.5.2 ABWE and Cepstral Truncation

The process of combining ABWE and cepstral truncation is performed in the following way. All the test files (adults' and children's) are first subjected to ABWE. The bandwidth extended files are segregated based on their warping factor values. The test files for each warping factor provide best performance for a specific truncation value. This is along the expected lines as each warping factor represents a specific vocal tract length and pitch value. Thus the best performance is taken for each warping factor case and then combined to obtain the final combination result. Table 3.6 gives the different experimental results for adults' test speech. The highlighted values in each column represents the best number for particular warping factor. These highlighted values are taken and combined to get the average value for adults' case which comes out to be 0.58% which is marginal improvement over 0.61%. This procedure is an optimistic approach in a theoritical sense to choose the best possible result. The practical implementation of the same needs an estimation method to choose the best possible result. The same is adopted in all the remaining results.

Table 3.7 gives the different experimental results for children's test speech. The highlighted values in each column represents the best number for particular warping factor. These highlighted values are taken and combined to get the average value for children's case which comes out to be 2.76% which is significant improvement over 4.01%.

3.5.3 ABWE, VTLN and Cepstral Truncation

The process of combining ABWE, VTLN and cepstral truncation is done in the following way. All the test files are subjected to ABWE at the first level. The bandwidth extended files are segregated based on their warping factor values. The segregated test files are warped using respective warping factors. The bandwidth extended and warped test files are subjected truncation. As in the earlier case, for each warping factor the best performance is achieved for a specific truncation value. But this truncation value will be different compared to the earlier case

Table 3.6: *Performance for the adults' test set on models trained on adults' speech data set for various truncations of base MFCC features along with their VTLN warp factor-wise breakup, Further respective utterences' MFCC are warped to respective breakup. The truncated base MFCC features are also appended with their corresponding first and second order temporal derivatives for recognition. The quantity in bracket gives the number of utterances corresponding to that VTLN warp factor.*

Base feature length	WER% for AD ABWE																
	VTLN Warping factor																
	0.80 (1)	0.82 (0)	0.84 (19)	0.86 (379)	0.88 (665)	0.90 (320)	0.92 (39)	0.94 (149)	0.96 (310)	0.98 (253)	1.00 (384)	1.02 (253)	1.04 (227)	1.06 (231)	1.08 (61)	1.10 (10)	1.12 (2)
16	0.00	–	0.00	0.28	0.38	0.37	0.00	0.66	0.00	0.51	0.38	0.21	1.02	0.76	2.67	0.00	0.00
15	0.00	–	0.00	0.38	0.52	0.46	0.00	0.66	0.00	0.51	0.53	0.11	0.90	0.63	2.67	0.00	0.00
14	0.00	–	0.00	0.56	0.52	0.46	0.00	0.66	0.00	0.64	0.53	0.21	0.79	0.76	2.67	0.00	0.00
13	0.00	–	0.00	0.56	0.75	0.37	0.00	0.44	0.00	**0.64**	0.76	**0.21**	**0.90**	1.01	2.67	0.00	0.00
12	0.00	–	0.00	**0.56**	0.70	0.46	0.00	0.44	0.10	1.03	0.76	0.32	1.02	1.14	2.67	0.00	0.00
11	0.00	–	0.00	0.66	0.66	0.46	0.00	**0.44**	0.10	1.03	0.76	0.54	1.02	1.14	2.67	0.00	0.00
10	0.00	–	0.00	0.75	0.66	0.55	0.00	0.66	0.00	0.90	0.76	0.54	0.90	1.01	2.67	0.00	0.00
9	0.00	–	0.00	0.85	0.66	0.46	0.00	0.66	**0.00**	0.90	0.76	0.64	1.02	1.01	2.67	0.00	0.00
8	0.00	–	0.00	1.03	0.70	0.37	0.00	0.88	0.20	0.77	0.76	0.64	1.02	**1.01**	2.67	**0.00**	0.00
7	0.00	–	**0.00**	1.03	0.66	**0.37**	0.00	0.66	0.20	0.90	0.68	0.64	0.90	1.39	**2.67**	16.67	**0.00**
6	0.00	–	2.63	0.94	**0.66**	0.64	0.00	0.66	0.40	1.03	**0.68**	0.64	1.02	1.52	3.21	19.11	40.00
5	0.00	–	7.89	1.03	1.08	0.73	**0.00**	0.66	0.60	1.29	1.07	0.54	1.24	2.15	3.21	16.67	60.00
4	0.00	–	5.26	1.41	1.50	1.10	1.79	2.19	0.79	1.93	1.60	0.86	2.49	3.03	4.28	11.11	20.00
3	**0.00**	–	15.79	3.67	4.08	4.02	7.14	6.58	3.08	3.99	4.19	3.01	4.41	6.69	5.88	75.00	75.00

due to the VTLN performed on the test files before truncation. Finally, the best performance is taken for each warping factor case and then combined to obtain the final combination result. Table 3.8 gives the different experimental results for adults' test speech. The highlighted values in each column represents the best number for particular warping factor. These highlighted values are taken and combined to get the average value for adults' case which comes out to be 0.50% which is marginal improvement over earlier results of 0.61% and 0.58%.

Table 3.9 tabulates the different experimental results for children's test speech. The highlighted values in each column represents the best number for particular warping factor. These highlighted values are taken and combined to get the average value for children's case which comes out to be 0.97% which is significant improvement over 4.01%, 2.76% and 1.06%. These results show the effectiveness of combining different approaches for children's speech recognition under mismatched condition.

Table 3.7: *Performance for the children's test set on models trained on adults' speech data set for various truncations of base MFCC features along with their VTLN warp factor-wise breakup. The truncated base MFCC features are also appended with their corresponding first and second order temporal derivatives for recognition. The quantity in bracket gives the number of utterances corresponding to that VTLN warp factor.*

Base feature length	WER% for CH ABWE																
	VTLN Warping factor																
	0.80 (22)	0.82 (1)	0.84 (510)	0.86 (1689)	0.88 (793)	0.90 (140)	0.92 (7)	0.94 (30)	0.96 (40)	0.98 (15)	1.00 (15)	1.02 (5)	1.04 (6)	1.06 (16)	1.08 (7)	1.10 (6)	1.12 (2)
16	8.57	0.00	14.05	5.86	2.32	2.40	0.00	0.00	4.11	0.00	0.00	0.00	0.00	1.45	0.00	0.00	0.00
15	8.57	0.00	12.30	5.40	2.35	2.16	0.00	0.00	4.11	0.00	0.00	0.00	0.00	1.45	0.00	0.00	0.00
14	5.71	0.00	10.27	4.79	2.17	2.16	0.00	0.00	2.74	0.00	0.00	0.00	0.00	2.90	0.00	0.00	0.00
13	5.71	0.00	8.45	4.20	2.10	1.44	0.00	0.00	4.11	0.00	0.00	0.00	0.00	2.90	0.00	0.00	0.00
12	5.71	0.00	7.70	3.87	2.10	1.44	0.00	**0.00**	5.48	0.00	0.00	0.00	0.00	2.90	0.00	0.00	0.00
11	5.71	0.00	7.23	3.74	1.89	1.92	0.00	0.91	6.85	0.00	0.00	0.00	0.00	2.90	0.00	0.00	0.00
10	5.71	0.00	6.28	3.46	1.85	2.40	0.00	0.91	6.85	0.00	0.00	0.00	0.00	2.90	0.00	0.00	0.00
9	5.71	0.00	5.81	3.58	**1.85**	2.40	0.00	0.00	4.11	0.00	0.00	0.00	0.00	2.90	0.00	0.00	0.00
8	8.57	0.00	4.93	**3.01**	1.96	2.16	0.00	0.91	5.48	0.00	0.00	0.00	0.00	1.45	0.00	0.00	0.00
7	5.71	0.00	5.34	3.28	2.07	2.40	0.00	0.91	4.11	1.96	2.86	0.00	0.00	**1.45**	0.00	0.00	0.00
6	5.71	0.00	5.61	3.94	2.35	1.68	0.00	0.91	4.11	3.92	0.00	0.00	0.00	2.90	3.70	0.00	0.00
5	5.71	0.00	**4.53**	3.88	2.57	**1.44**	0.00	1.82	4.11	0.00	**0.00**	0.00	0.00	4.35	3.70	0.00	14.29
4	**2.86**	0.00	5.81	5.25	3.67	1.68	0.00	0.91	**2.74**	**0.00**	5.71	0.00	**0.00**	4.35	**0.00**	0.00	**0.00**
3	11.43	**0.00**	11.22	8.14	5.74	5.77	**0.00**	3.64	4.11	3.92	5.71	**0.00**	6.67	14.49	7.41	**0.00**	28.57

3.6 Summary

The chapter started with the motivation for the work and it was demonstrating the significance of ABWE for speech recognition. The first study described the development of TIDIGITS based digit recognition system using MFCC and HMM. For comparison purpose, ASR system is developed for both adults' and children's speech and also in each, narrowband and wideband cases. As a practical use, since most of the data is available for adults' speech case, the models are trained using adults' speech and then tested using either adults' or children's speech. When narrowband speech trained adults' speech models are tested with narrowband adults' speech, the performance is best demonstrating the matching condition. Further, the wideband testing case of adults' speech further improves performance compared to the narrowband slightly in case of adults' speech inferring that there is not much high frequency information in case of adults.

The narrowband children's speech tested against narrowband adults' speech models gave the worst performance indicating the significant mismatch. Alternatively, the wideband chil-

Table 3.8: *Performance for the adults' test set on models trained on adults' speech data set for various truncations of base MFCC features along with their VTLN warp factor-wise breakup. Further respective utterances MFCCs are warped using respective warp factor. The truncated base MFCC features are also appended with their corresponding first and second order temporal derivatives for recognition. The quantity in bracket gives the number of utterances corresponding to that VTLN warp factor.*

Base feature length	WER% for AD ABWE + VTLN																
	VTLN Warping factor																
	0.80 (1)	0.82 (0)	0.84 (19)	0.86 (379)	0.88 (665)	0.90 (320)	0.92 (39)	0.94 (149)	0.96 (310)	0.98 (253)	1.00 (384)	1.02 (253)	1.04 (227)	1.06 (231)	1.08 (61)	1.10 (10)	1.12 (2)
16	0.00	–	0.00	0.38	0.42	0.27	0.00	0.88	0.00	0.51	0.38	0.11	0.90	1.01	2.67	0.00	0.00
15	0.00	–	0.00	0.38	0.47	0.27	0.00	0.88	0.00	0.51	0.53	0.11	0.79	1.01	2.67	0.00	0.00
14	0.00	–	0.00	0.38	0.56	0.37	0.00	0.66	0.00	0.64	0.53	**0.21**	0.90	0.88	2.67	0.00	0.00
13	0.00	–	0.00	0.38	0.56	0.37	0.00	0.66	0.00	**0.64**	0.76	0.21	0.68	**1.01**	2.67	0.00	0.00
12	0.00	–	0.00	0.38	0.52	0.27	0.00	0.66	0.00	1.03	0.76	0.32	**0.68**	1.14	2.67	0.00	0.00
11	0.00	–	0.00	0.38	0.52	0.37	0.00	0.66	0.00	1.03	0.76	0.43	1.02	1.14	2.67	0.00	0.00
10	0.00	–	0.00	**0.38**	**0.47**	0.37	0.00	0.44	**0.00**	0.77	0.76	0.54	0.90	1.14	2.67	0.00	0.00
9	0.00	–	0.00	0.47	0.61	0.37	0.00	**0.44**	0.10	0.90	0.76	0.64	1.13	1.14	2.67	0.00	0.00
8	0.00	–	0.00	0.47	0.66	**0.27**	0.00	0.66	0.20	0.77	0.76	0.64	1.02	1.14	2.67	0.00	0.00
7	0.00	–	**0.00**	0.85	0.75	0.46	0.00	0.66	0.30	0.90	0.68	0.64	0.79	1.26	**2.67**	0.00	0.00
6	0.00	–	2.63	0.75	0.94	0.55	0.00	0.66	0.40	1.29	**0.68**	0.64	0.90	1.26	3.21	**0.00**	0.00
5	0.00	–	2.63	1.03	1.17	0.82	**0.00**	1.10	0.60	1.29	1.07	0.64	0.90	1.64	3.21	16.67	**0.00**
4	0.00	–	2.63	1.13	1.41	1.19	1.79	1.97	1.09	2.19	1.60	0.86	1.81	2.40	4.28	11.11	40.00
3	**0.00**	–	7.89	3.01	4.17	4.48	7.14	6.80	3.17	3.99	4.19	2.79	3.85	4.92	5.88	11.11	40.00

dren's speech tested against wideband adults' speech models gave significant improvement in performance compared to narrowband case inferring the presence of significant high frequency information for children's case. This also indicates that, if ABWE method is used then, we may benefit in children's speech case. To experimentally verify this fact, an existing mostly used ABWE method is implemented and ASR study is performed. The ABWE extended children's speech provided significant performance improvement compared to its narrowband case, demonstrating the significance of ABWE for children's speech recognition under mismatched condition. However, ABWE seem to be not effective for adults' speech as there is not much information in the high frequency range.

The VTLN alone also improved the ASR performance significantly in case of children's speech, for both narrowband and wideband cases. This will tempt to conclude that VTLN will suffice to reduce mismatch in case of children's speech. However, the combination of VTLN with ABWE further improves performance as compared any of the individual cases. This shows that there is some different information in both VTLN and ABWE methods. This is true also,

Table 3.9: *Performance for the children's test set on models trained on adults' speech data set for various truncations of base MFCC features along with their VTLN warp factor-wise breakup, Further respective utterances' MFCC are warped to respective breakup. The truncated base MFCC features are also appended with their corresponding first and second order temporal derivatives for recognition. The quantity in bracket gives the number of utterances corresponding to that VTLN warp factor.*

Base feature length	WER% for CH ABWE + VTLN																
	VTLN Warping factor																
	0.80 (22)	0.82 (1)	0.84 (510)	0.86 (1689)	0.88 (793)	0.90 (140)	0.92 (7)	0.94 (30)	0.96 (40)	0.98 (15)	1.00 (15)	1.02 (5)	1.04 (6)	1.06 (16)	1.08 (7)	1.10 (6)	1.12 (2)
16	2.86	0.00	1.01	1.10	0.86	0.72	0.00	0.00	5.48	0.00	0.00	0.00	0.00	1.45	0.00	0.00	0.00
15	2.86	0.00	1.15	1.02	**0.86**	0.48	0.00	0.00	5.48	0.00	0.00	0.00	0.00	1.45	0.00	0.00	0.00
14	5.71	0.00	0.95	1.05	0.93	0.96	0.00	0.00	4.11	0.00	0.00	0.00	0.00	1.45	0.00	0.00	0.00
13	5.71	0.00	0.95	1.10	0.96	0.96	0.00	0.91	4.11	0.00	0.00	0.00	0.00	1.45	0.00	0.00	0.00
12	2.86	0.00	**0.88**	1.05	**0.96**	0.48	0.00	0.91	5.48	0.00	0.00	0.00	0.00	1.45	0.00	0.00	0.00
11	2.86	0.00	1.01	**1.05**	1.07	0.72	0.00	0.00	6.85	0.00	0.00	0.00	0.00	1.45	0.00	0.00	0.00
10	2.86	0.00	0.95	1.23	1.07	0.72	0.00	0.00	5.48	0.00	0.00	0.00	0.00	2.90	0.00	0.00	0.00
9	2.86	0.00	1.22	1.32	1.18	**0.72**	0.00	0.00	5.48	0.00	0.00	0.00	0.00	2.90	0.00	0.00	0.00
8	2.86	0.00	1.28	1.25	1.21	1.20	0.00	0.00	4.11	0.00	0.00	0.00	0.00	1.45	0.00	0.00	0.00
7	2.86	0.00	1.49	1.37	1.07	0.96	0.00	**0.00**	4.11	1.96	2.86	0.00	0.00	1.45	0.00	0.00	0.00
6	**2.86**	0.00	1.76	2.03	1.82	1.68	0.00	0.91	2.74	3.92	0.00	0.00	0.00	**1.45**	0.00	0.00	0.00
5	5.71	0.00	2.03	2.65	2.00	1.44	0.00	1.82	4.11	0.00	**0.00**	0.00	0.00	2.90	3.70	0.00	0.00
4	8.57	0.00	3.51	3.81	2.96	1.68	0.00	0.91	**2.74**	**0.00**	5.71	0.00	**0.00**	4.35	3.70	0.00	**0.00**
3	8.58	**0.00**	8.45	7.61	6.13	5.53	**0.00**	3.64	4.11	3.92	5.71	**0.00**	6.67	11.59	**0.00**	**0.00**	28.57

as VTLN only performs normalization only within the given band, where as, ABWE tries to reconstruct information outside the band also.

The truncation of MFCC coefficients also provides significant performance improvement in children's speech case and degradation in case of adults' speech. This shows that the pitch and vocal tract length mismatch is significantly high in case of children's speech. Alternatively, truncation leads to removal of crucial vocal tract information in case of adults' speech. Thus truncation of coefficients is effective in case of children's speech recognition under mismatched condition. Compared to narrowband, the wideband case provides better performance for children's speech in the truncated features case. This shows that, higher band information helps in children's speech recognition along with cepstral truncation. Therefore using ABWE along with cepstral truncation provided a performance which is significantly better compared to only by truncation of coefficients. Thus the information offered by both the cases are different. The combination of VTLN, ABWE and truncation approaches provides the best performance.

The following conclusions can be made from the studies performed in this chapter: (1)

ABWE is effective for children's speech recognition under mismatched condition. (2) The information offered by ABWE approach is different compared to those of VTLN and cepstral truncation and can be used in combination with any of these methods for further improving performance. (3) The combination study indeed infers this fact. (4) Therefore development any new ABWE method is an added value to the children's speech recognition under mismatched condition.

4

Proposed ABWE Improvements using Auxiliary Information

Contents

The previous chapter demonstrated the significance of ABWE for children's speech recognition under mismatched condition. The ABWE indeed improves the ASR performance significantly. Also, the information provided by ABWE is different compared to that of VTLN or coefficient truncation which are earlier used for children's speech recognition under mismatched condition. As a result any new ABWE method developed will be an added value to VTLN and coefficient truncation approaches. The motivation is therefore to develop new methods for ABWE and demonstrate their usefulness for children's speech recognition under mismatched condition.

There are several methods in the literature for ABWE [10–14]. However, most of them are based on source-system or vocoder framework. In the vocoder framework, the ABWE speech is synthesized from the extended vocal tract and excitation source information. Also, the improvement achieved is measured in terms of improvement in the perceptual quality. In the present work, the focus is on improving the ASR performance and for this, the ABWE features that provide improved ASR performance will suffice without worrying about the synthesis part and also resulting speech perceptual quality. Also, it may be noted that improving perceptual quality may not necessarily improve the ASR performance. Therefore the focus of this work is to develop some ABWE methods that improve children's speech recognition under mismatched condition.

As described in the previous chapter, the framework considered for the study is an availability of ASR system where the models are trained using speech from adult speakers. This is a practical assumption as most of the speech data available for building ASR system is mostly collected from adult speakers. In such a scenario, if the system is to be used by children, then due to mismatch, the ASR performance decreases significantly. To improve the performance, the mismatch needs to be reduced. The VTLN only adapts the available information in the given children's speech data into the adults' speech model based on vocal tract length normalization. The coefficient truncation smooths the fluctuations present in the short term spectra and accordingly reduces the mismatch. In both these no new information is added to the available children's speech. However, as analyzed in the previous chapters, the children's speech

will have most of the information in the high frequency range and the narrowband version of children's speech has lot of information missing in it. Therefore the most beneficial one is to perform ABWE on the given NB children's speech.

The first difference that will happen between features of speech from adult and child speakers is the variation in the distribution of the feature space. To keep the variabilities minimum, a suggested approach is to use class-specific information. That is constrain the training and testing process using the class-specific knowledge. At the gross-level, there is significant difference between adults' and children's speech. Due to rapid change in the vocal tract and source physical structure and associated dynamics, the variability within the broad category children's speech may be very high compared to adults' speech. As a result benefit may be achieved by exploiting the age-specific information in case of children's speech. Since the dynamics associated with speech production depends on the age of the speaker, the same can be exploited by computing the dynamic features and using them along with standard static features.

All the above mentioned observations represent different aspects and children's speech and hence are termed as *auxiliary information*. This chapter develops ABWE methods by exploiting this auxiliary information. The rest of the chapter is organized as follows: Section 4.1 describes different statistical measures and also computing them for different cases of auxiliary information to motivate the present work. The development of ABWE method using class-specific information is described in Section 4.2. Section 4.3 describes development of feature domain MFCC based ABWE. The ABWE using age-specific and delta features is described in Section 4.4. The chapter is summarized and concluded in Section 4.5.

4.1 Comparison of Children's and Adults' Speech using Statistical Measures

The dependency between NB and HB exists due to the physical nature of human speech production. Further, this dependency is affected by the speaker's vocal tract configuration. Accordingly, we can expect the variability in the dependency between NB and HB depending on whether the speech is from children or adult. Thus the dependency information is useful for

developing new ABWE methods.

We can investigate and quantify the dependency between the spectral envelopes of speech in disjoint frequency bands, namely, NB and HB using the concept of mutual information. Let $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_M\}$ and $\mathbf{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_M\}$ be the feature sequences corresponding to NB and HB, respectively.

4.1.1 Mutual Information (I)

The mutual information between the NB and HB feature sequences is computed as

$$\widehat{I(\mathbf{X};\mathbf{Y})} = \frac{1}{M}\sum_{m=1}^{M} log_2\left(\frac{f_{GMM}(\mathbf{X}_m, \mathbf{Y}_m)}{f_{GMM}(\mathbf{X}_m) f_{GMM}(\mathbf{Y}_m)}\right) \tag{4.1}$$

where $f_{GMM}(.)$ is the GMM density function and $\widehat{I(.;.)}$ is the estimate of the mutual information. $f_{GMM}(\mathbf{X}_m)$ and $f_{GMM}(\mathbf{Y}_m)$ are calculated as the marginal distribution from the joint distribution $f_{GMM}(\mathbf{X}_m, \mathbf{Y}_m)$.

4.1.2 Differential Entropy (H)

The entropy can be used to measure the amount of new information or degree of uncertainty present in the given feature sequence. The differential entropy of the variable Y can be defined as

$$h(Y) = -\int_{\Omega_y} f_Y(y) log_2(f_Y(y)) dy \tag{4.2}$$

where Ω_y is the value space of Y.

Replacing the integral with a summation and the pdf by the corresponding probability mass function, we obtain the definition of entropy, $H(Y)$. Furthter, in order to compute the mutual information between the high-band, the GMM models describing the joint and marginal pdfs are used. A direct computation of the mutual information or differential entropy from the pdf's modeled by the GMM is non-trivial for models with more than one mixture component. Thus the numnerical method of stochastic integration is used. Based on these approximations, the

differential entropy of the HB is estimated as

$$\widehat{H(\mathbf{Y})} = -\frac{1}{M}\sum_{m=1}^{M} log_2\left(f_{GMM}(\mathbf{Y}_m)\right) \tag{4.3}$$

4.1.3 Ratio Measure (R_{IH})

Given the NB and HB, the mutual information gives the dependency among the two disjoint bands and the differential entropy gives amount of new information present in the HB. For ABWE, our interest is in finding how large the remaining uncertainty of the HB is, given the NB. This is done by determining the ratio between the mutual information of the two bands and the entropy of the HB given by

$$\widehat{R}_{IH}(\%) = \frac{\widehat{I(\mathbf{X};\mathbf{Y})}}{\widehat{H(\mathbf{Y})}} \times 100 \tag{4.4}$$

4.1.4 Separability (ε)

The separability measure is well known from statistics [138]. It quantifies the discriminative power of a feature set for a classification task. The separability is known as a measure for the quality of a particular feature set for a classification problem [138]. The separability measure can be calculated from a labelled set of training data, i.e. for each feature vector in the set, the corresponding class must be known. Let Ξ denote the set of feature vectors $\mathbf{x}$ assigned to the i-th class. The number of feature vectors in the i-th set is $N_{\Xi_i} = |\Xi_i|$. Let N_s be the total number of classes for the classification task and let N_m be the total number of feature vectors in the training data (from all classes).

From the training data, the within-class covariance matrix is given by

$$\mathbf{V}_{\mathbf{x}} = \frac{1}{N_m}\sum_{i=1}^{N_s}\sum_{\mathbf{x}\in\Xi_i}(\mathbf{x}-\mu_i)(\mathbf{x}-\mu_i)^T \tag{4.5}$$

and the between-class covariance matrix is given by

$$\mathbf{B}_{\mathbf{x}} = \sum_{i=1}^{N_s}\frac{N_{\Xi_i}}{N_m}(\mu_i-\mu)(\mu_i-\mu)^T \tag{4.6}$$

are calculated, where

$$\mu_i = \frac{1}{N_{\Xi_i}} \sum_{\mathbf{x} \in \Xi_i} \mathbf{x} \text{ and } \mu = \sum_{i=1}^{N_s} \frac{N_{\Xi_i}}{N_m} \mu_i \tag{4.7}$$

The separability measure shall be larger if the between-class covariance gets smaller or if the within-class covariance gets larger. Accordingly, the separability measure is empirically defined by the term

$$J_{\mathbf{x}} = \mathbf{V}_{\mathbf{x}}^{-1} \mathbf{B}_{\mathbf{x}} \tag{4.8}$$

To obtain a scalar measure for the separability of the classes a trace criterion is used [138]

$$\varepsilon(\mathbf{x}) = \text{tr}(J_{\mathbf{x}}) = \text{tr}(\mathbf{V}_{\mathbf{x}}^{-1} \mathbf{B}_{\mathbf{x}}) \tag{4.9}$$

The separability depends on the definition of the classes. Comparing $\varepsilon(\mathbf{x})$ for different feature vectors $\mathbf{x}$ with the same class definitions, a larger value indicates a better suitability of the corresponding feature vector for classification and estimation

Table 4.1 lists the mutual information between NB and HB, HB entropy and ratio measure computed for different classes (digits) for speech of both children and adults. These measures were applied over the speech signals after different types of ABWE transformation (global and class specific). The transformations are learnt using the training data, different from that of test data. To compute the measures, the test data was applied on the learnt transformations. It can be observed that the I, H and R_{IH} varies across different classes. It is to note that the averaged R_{IH} increases from 3.03% for global transformation case to 11.64% using class specific transformation. This demonstrates the significance of exploiting class-specific information for ABWE. It is also interesting to note that the increase in the averaged R_{IH} for children's speech is less than that for adults speech. This trend may be attributed to the loss of spectral information for children in case of narrowband and and also higher variability in the vocal tract length. The separability measure shows a lower value for the children's speech compared to that of adult. This may be attributed to significant overlap of class-specific information in the feature space. However, since R_{IH} shows significant increase, ABWE method can be developed using class (digit) specific information.

Table 4.1: *The mutual information ($\widehat{I(X;Y)}$), high-band entropy ($\widehat{H(Y)}$), and their ratio ($\widehat{R}_{IH}$) for children's and adults' speech with application of global and class specific ABWE transform.Separability $\varepsilon(\mathbf{x})$ for children's and adults' speech with application of global ABWE transform.*

Class	Children					
	Global transformation			Class specific transformation		
	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %
one	0.96	47.65	2.02	5.35	50.46	10.61
two	1.66	48.77	3.40	6.09	50.50	12.06
three	1.39	48.74	2.85	4.81	49.47	9.72
four	1.25	48.05	2.60	5.01	50.64	9.89
five	1.57	47.52	3.31	8.65	52.89	16.36
six	3.05	50.02	6.09	5.41	50.10	10.60
seven	1.86	48.67	3.83	6.71	50.55	13.28
eight	2.26	48.31	4.69	6.77	51.50	13.15
nine	0.75	47.86	1.58	6.10	50.35	12.11
zero	0.49	48.99	1.00	3.59	47.97	7.48
oh	0.86	48.33	1.78	6.48	50.23	12.89
Avg.	1.47	48.47	3.03	5.90	50.50	11.64
Class	Adults					
	Global transformation			Class specific transformation		
	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %
one	0.86	48.78	1.76	6.23	51.94	11.99
two	1.13	48.27	2.34	6.78	52.10	13.02
three	1.05	47.99	2.20	6.23	51.69	12.07
four	1.60	47.48	3.36	6.48	52.69	12.30
five	1.98	47.86	4.13	9.51	52.80	18.02
six	3.14	50.42	6.22	7.43	51.85	14.32
seven	1.62	48.91	3.31	9.59	52.17	18.39
eight	1.32	48.46	2.73	7.03	51.63	13.61
nine	0.97	47.53	2.04	7.82	51.70	15.13
zero	0.84	48.61	1.73	3.71	48.98	7.57
oh	1.50	47.95	3.12	6.59	51.17	12.88
Avg.	1.48	48.42	3.05	7.05	51.68	13.59

	Separability $\varepsilon(\mathbf{x})$	
	Children	Adults
	Global transformation	Global transformation
Avg.	4.70	6.22

Table 4.2 shows the I, H, R_{IH} and ε computed by exploiting the age-specific information. It can be noted that, there is no increase in the R_{IH} value by exploiting the age-specific information. However, the separability value increases which demonstrates that, the feature set using age-specific information is more discriminative compared to global transform. Therefore age-specific information can also be exploited for ABWE.

Table 4.2: *The mutual information ($\widehat{I(X;Y)}$), the high-band entropies ($\widehat{H(Y)}$), and their ratios ($\widehat{R}_{IH}$) for children's speech with application of global and age specific ABWE transforms. Separability $\varepsilon(\mathbf{x})$ for children's speech with application of global ABWE and age specific transform.*

Children						
Age in years	Global transformation			Age Specific transform		
	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %
06	7.43	32.85	22.61	7.44	33.19	22.43
07	6.05	32.76	18.47	5.92	32.89	18.01
08	6.12	32.78	18.68	6.06	32.99	18.38
09	6.05	32.84	18.41	5.97	33.10	18.04
10	5.83	32.89	17.73	5.74	33.12	17.33
11	5.82	32.76	17.76	5.70	33.00	17.27
12	5.46	32.85	16.61	4.93	33.00	14.95
13	5.19	32.79	15.83	4.31	32.84	13.14
14	5.22	32.86	15.89	4.29	32.88	13.04
15	5.14	32.80	15.67	4.24	32.82	12.91
Avg.	5.87	32.81	17.88	5.68	33.02	17.20
	Separability $\varepsilon(\mathbf{x})$					
	Global transformation			Age Specific transform		
Avg.	7.62			8.18		

Table 4.3 shows that I, H, R_{IH} and ε values computed using the delta (Δ) features. The delta features, essentially refers to the measure of change happening in the feature sequence. That is how, the features are changing with time. Using delta features, increases both the ratio as well as separability values inferring that the delta (Δ) features can be exploited for ABWE.

Table 4.3: *The mutual information ($\widehat{I(X;Y)}$), the high-band entropies ($\widehat{H(Y)}$), and their ratios ($\widehat{R}_{IH}$) for children's speech with application of global and age specific ABWE transforms with Δ. Separability $\varepsilon(\mathbf{x})$ for children's speech with application of global ABWE and age specific transform with Δ.Half window size, $\Theta = 13$ is selected to compute Δ.*

Children						
Age in years	Global transformation with Δ			Age Specific transform with Δ		
	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %
06	16.37	47.01	34.83	27.36	51.64	52.98
07	12.79	45.10	28.36	15.21	44.29	34.33
08	12.92	45.49	28.39	15.38	44.41	34.62
09	13.21	45.72	28.90	14.05	43.42	32.35
10	12.62	45.42	27.79	14.28	43.48	32.84
11	12.46	45.06	27.65	14.75	44.20	33.37
12	12.12	45.27	26.78	15.06	44.37	33.94
13	11.91	45.28	26.31	15.01	44.49	33.74
14	11.95	45.38	26.33	15.19	44.67	34.00
15	11.81	45.21	26.13	15.15	44.61	33.97
Avg.	12.68	45.36	27.96	14.92	44.14	33.80

	Separability $\varepsilon(\mathbf{x})$	
	Global transformation with Δ	Age Specific transform with Δ
Avg.	7.86	8.57

Table 4.4 shows the I, H and R_{IH} for different values of Δ computed in the global transformation case. In this case, no age information is used. As it can be observed, the computed values are sensitive to the Δ value. That is, as the Δ value increases, the values of all the parameters increases. This shows the significance of Δ feature.

Table 4.4: *The mutual information ($\widehat{I(X;Y)}$), the high-band entropies ($\widehat{H(Y)}$), and their ratios ($\widehat{R}_{IH}$) for children's speech with application of global ABWE transforms with Δ.Half window size, Θ is selected between range of 1 to 15 to compute Δ.*

ABWE-GT+Δ															
Age in years	Half window size, Θ														
	1			2			3			4			5		
	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %
06	11.66	34.87	33.44	12.31	37.99	32.40	12.26	39.13	31.32	13.34	41.35	32.27	13.05	42.44	30.75
07	10.42	34.40	30.29	11.11	37.44	29.66	10.87	38.53	28.20	11.65	40.51	28.76	11.34	41.43	27.38
08	10.48	34.70	30.19	11.46	37.81	30.30	11.05	38.87	28.42	11.93	40.89	29.18	11.65	41.83	27.84
09	10.54	34.71	30.38	11.59	37.88	30.60	11.15	38.92	28.64	12.04	40.97	29.40	11.72	41.92	27.95
10	10.37	34.63	29.96	11.25	37.81	29.74	10.81	38.85	27.82	11.69	40.88	28.59	11.42	41.81	27.32
11	10.21	34.32	29.75	11.15	37.43	29.78	10.67	38.50	27.71	11.36	40.44	28.10	11.18	41.40	27.01
12	10.06	34.61	29.06	11.01	37.68	29.23	10.45	38.79	26.95	11.31	40.69	27.79	11.06	41.64	26.56
13	9.93	34.64	28.69	10.90	37.69	28.92	10.29	38.80	26.53	11.15	40.69	27.41	10.88	41.63	26.13
14	9.93	34.69	28.63	10.95	37.76	28.99	10.32	38.86	26.56	11.23	40.77	27.53	10.93	41.71	26.20
15	9.90	34.61	28.61	10.86	37.66	28.82	10.26	38.78	26.47	11.12	40.65	27.35	10.84	41.59	26.07
Avg.	10.33	34.56	29.90	11.27	37.68	29.90	10.82	38.74	27.92	11.64	40.73	28.59	11.38	41.67	27.32
Age in years	Half window size, Θ														
	6			7			8			9			10		
	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %
06	13.45	43.29	31.08	13.76	44.22	31.11	13.68	44.34	30.86	13.47	44.96	29.96	14.55	45.58	31.91
07	11.63	42.23	27.53	11.67	42.95	27.17	11.61	43.13	26.92	11.29	43.66	25.86	12.10	44.21	27.36
08	12.12	42.64	28.41	12.15	43.40	27.99	11.97	43.54	27.49	11.62	44.04	26.38	12.39	44.57	27.79
09	12.15	42.72	28.45	12.14	43.46	27.93	12.00	43.60	27.53	11.74	44.14	26.60	12.67	44.74	28.32
10	11.76	42.61	27.61	11.71	43.33	27.03	11.55	43.45	26.57	11.27	43.98	25.62	12.11	44.53	27.19
11	11.66	42.20	27.63	11.59	42.92	27.01	11.47	43.11	26.61	11.18	43.62	25.63	12.02	44.23	27.19
12	11.44	42.46	26.96	11.27	43.14	26.12	11.23	43.33	25.91	10.87	43.81	24.81	11.76	44.41	26.48
13	11.30	42.45	26.62	11.08	43.14	25.69	11.06	43.32	25.52	10.70	43.80	24.42	11.57	44.40	26.05
14	11.41	42.54	26.83	11.16	43.22	25.83	11.13	43.40	25.64	10.79	43.89	24.58	11.63	44.49	26.15
15	11.23	42.41	26.49	11.00	43.09	25.53	11.00	43.27	25.42	10.62	43.75	24.27	11.53	44.35	26.00
Avg.	11.80	42.48	27.79	11.75	43.20	27.21	11.64	43.37	26.83	11.33	43.88	25.82	12.18	44.46	27.4
Age in years	Half window size, Θ														
	11			12			13			14			15		
	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %
06	14.03	46.06	30.45	15.27	46.56	32.81	**16.37**	47.01	**34.83**	15.91	47.49	33.50	15.55	47.18	32.97
07	11.59	44.62	25.97	12.46	44.88	27.77	**12.79**	45.10	**28.36**	12.35	45.52	27.13	12.22	45.51	26.86
08	11.91	44.98	26.48	12.60	45.25	27.85	**12.92**	45.49	**28.39**	12.49	45.93	27.19	12.37	45.92	26.95
09	12.19	45.19	26.98	12.85	45.45	28.27	**13.21**	45.72	**28.90**	12.87	46.19	27.86	12.72	46.17	27.55
10	11.62	44.92	25.86	12.24	45.18	27.10	**12.62**	45.42	**27.79**	12.28	45.87	26.77	12.20	45.89	26.58
11	11.43	44.60	25.64	12.04	44.84	26.84	**12.46**	45.06	**27.65**	12.11	45.50	26.61	11.99	45.54	26.34
12	11.24	44.79	25.09	11.72	45.02	26.03	**12.12**	45.27	**26.78**	11.85	45.71	25.93	11.80	45.75	25.79
13	11.06	44.79	24.70	11.52	45.02	25.59	**11.91**	45.28	**26.31**	11.68	45.72	25.55	11.62	45.77	25.40
14	11.12	44.88	24.77	11.56	45.11	25.63	**11.95**	45.38	**26.33**	11.73	45.82	25.60	11.66	45.87	25.41
15	11.02	44.74	24.63	11.44	44.95	25.45	**11.81**	45.21	**26.13**	11.58	45.65	25.36	11.52	45.69	25.21
Avg.	11.66	44.86	25.99	12.30	45.11	27.26	**12.68**	45.36	**27.96**	12.33	45.80	26.91	12.22	45.81	26.67

Table 4.5 shows I, H and R_{IH} for different values of Δ computed in the age-specific case. As in the global transform case, it can be observed that the computed values are sensitive to the Δ value. That is, as the Δ value increases, the values of all the parameters increases. However, the use of age-specific information reflects the speaking rate information represented in the form of higher values for these parameters. In case of children with smaller age, the values of the parameters are high for higher delta values. Alternatively, the values are higher in case of higher age groups children for smaller values of Δ. This infers that, the variability associated with children's speaking rate is indeed reflected in the estimated parameters of I, H and R_{IH} using age-specific information.

Table 4.5: *The mutual information ($\widehat{I(X;Y)}$), the high-band entropies ($\widehat{H(Y)}$), and their ratios ($\widehat{R}_{IH}$) for children's speech with application of age Specific ABWE transforms with Δ.Half window size, Θ is selected between range of 1 to 15 to compute Δ.*

ABWE-AG+Δ															
Age in years	Half window size, Θ														
	1			2			3			4			5		
	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %
06	18.68	39.13	47.74	18.77	42.64	44.02	21.63	45.38	47.67	23.08	47.21	48.89	24.12	48.58	49.66
07	11.44	35.07	32.61	12.29	38.27	32.12	12.80	39.67	32.26	13.91	41.35	33.64	14.69	42.14	34.86
08	12.06	34.82	34.63	12.54	37.02	33.87	12.76	38.59	33.08	13.14	40.22	32.67	14.19	41.70	34.03
09	12.07	34.06	35.44	12.11	37.32	32.44	12.73	38.57	33.01	13.08	40.27	32.49	12.99	40.93	31.74
10	10.97	34.21	32.09	12.27	36.89	33.26	12.46	38.28	32.55	12.88	39.98	32.22	13.94	41.20	**33.84**
11	12.38	35.14	35.22	12.45	38.16	32.62	13.51	39.95	33.81	13.71	41.23	33.25	14.77	42.39	34.86
12	12.29	35.03	**35.09**	12.58	38.07	33.06	13.80	39.96	34.54	13.93	41.07	33.92	14.47	42.18	34.31
13	12.43	35.19	**35.31**	12.81	38.18	33.56	13.88	40.16	34.56	14.23	41.39	34.39	14.75	42.45	34.74
14	12.55	35.26	**35.60**	12.91	38.29	33.72	14.00	40.27	34.76	14.25	41.46	34.38	14.85	42.57	34.89
15	12.41	35.26	**35.21**	12.78	38.29	33.37	13.95	40.21	34.69	14.24	41.43	34.38	14.77	42.55	34.71
Avg.	12.06	34.81	34.65	12.48	37.70	33.11	13.20	39.31	33.59	13.58	40.80	33.28	14.34	41.90	34.22
Age in years	Half window size, Θ														
	6			7			8			9			10		
	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %
06	25.13	49.75	50.51	26.35	50.62	52.06	27.36	51.64	52.98	27.99	52.53	53.28	26.58	52.51	50.63
07	15.01	43.02	34.88	15.56	43.74	**35.57**	15.21	44.29	34.33	15.01	44.59	33.66	14.90	45.11	33.04
08	14.77	42.79	34.52	15.29	43.71	34.98	15.38	44.41	34.62	15.56	44.95	34.62	15.43	45.21	34.13
09	13.80	42.19	32.72	14.20	42.87	**33.12**	14.05	43.42	32.35	14.19	43.95	32.29	14.28	44.44	32.14
10	14.11	42.03	33.58	14.30	42.92	33.32	14.28	43.48	32.84	14.13	43.87	32.21	13.60	44.09	30.85
11	15.42	43.40	35.52	14.97	43.73	34.23	14.75	44.20	33.37	15.24	44.83	33.99	15.97	45.45	35.14
12	14.20	42.88	33.11	14.44	43.55	33.16	15.06	44.37	33.94	15.16	44.82	33.84	14.90	45.06	33.06
13	14.34	43.05	33.31	14.62	43.76	33.40	15.01	44.49	33.74	15.16	44.96	33.73	15.11	45.28	33.36
14	14.54	43.21	33.65	14.75	43.92	33.59	15.19	44.67	34.00	15.33	45.13	33.97	15.31	45.45	33.68
15	14.45	43.17	33.48	14.71	43.86	33.54	15.15	44.61	33.97	15.38	45.12	34.09	15.26	45.42	33.59
Avg.	14.74	42.87	34.38	14.91	43.54	34.24	14.92	44.14	33.80	15.09	44.65	33.80	15.13	45.05	33.58
Age in years	Half window size, Θ														
	11			12			13			14			15		
	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %	$\widehat{I(X;Y)}$ [bits]	$\widehat{H(Y)}$ [bits]	$\widehat{R}_{IH}$ %
06	28.83	53.74	53.64	27.83	53.77	51.75	27.97	54.13	51.67	30.18	55.25	54.62	**30.67**	55.61	**55.16**
07	14.71	45.61	32.25	14.98	46.19	32.42	15.05	46.50	32.36	15.72	46.99	33.46	**16.44**	47.67	34.49
08	15.79	45.72	34.53	15.92	46.17	34.47	16.09	46.61	34.53	**16.35**	47.04	**34.76**	15.88	47.47	33.46
09	13.91	44.82	31.03	14.92	45.59	32.73	15.16	45.87	33.04	15.31	46.25	33.11	**15.45**	46.69	33.10
10	13.49	44.46	30.34	14.05	45.30	31.03	13.44	45.28	29.68	13.67	45.71	29.90	**13.82**	46.12	29.97
11	**16.35**	45.87	**35.64**	15.44	46.09	33.49	15.79	46.36	34.07	14.94	46.56	32.09	15.12	46.95	32.21
12	**15.19**	45.44	33.43	14.88	45.69	32.57	14.71	46.12	31.90	14.29	46.34	30.84	14.34	46.71	30.70
13	**15.41**	45.66	33.75	15.05	45.86	32.81	14.99	46.30	32.37	14.57	46.53	31.32	14.72	46.97	31.34
14	**15.51**	45.79	33.88	15.21	46.03	33.04	15.16	46.46	32.62	14.75	46.70	31.59	14.86	47.10	31.55
15	**15.48**	45.78	33.82	15.15	46.00	32.93	15.10	46.44	32.51	14.70	46.70	31.47	14.80	47.10	31.43
Avg.	15.27	45.48	33.57	15.26	45.95	33.20	15.31	46.24	33.11	15.20	46.58	32.63	15.29	47.01	32.53

Motivated by earlier work [90], we also calculated the mutual information between bands with inclusion of Δ by combining static and delta features as $[C_0 - C_4, \Delta C_0 - \Delta C_4]$ and $[C_0 - C_2, \Delta C_0 - \Delta C_2]$ for LB and HB, respectively. Thus the length of features is kept identical for with/without delta features. Fig. 4.1 shows the plots for both global and age-specific ABWE cases for with/without delta features. It is to note that without deltas in MFCC feature, the ratio for global and age-specific ABWE are very close. However, the relative increase is much more for age-specific ABWE than that for global ABWE with delta features.

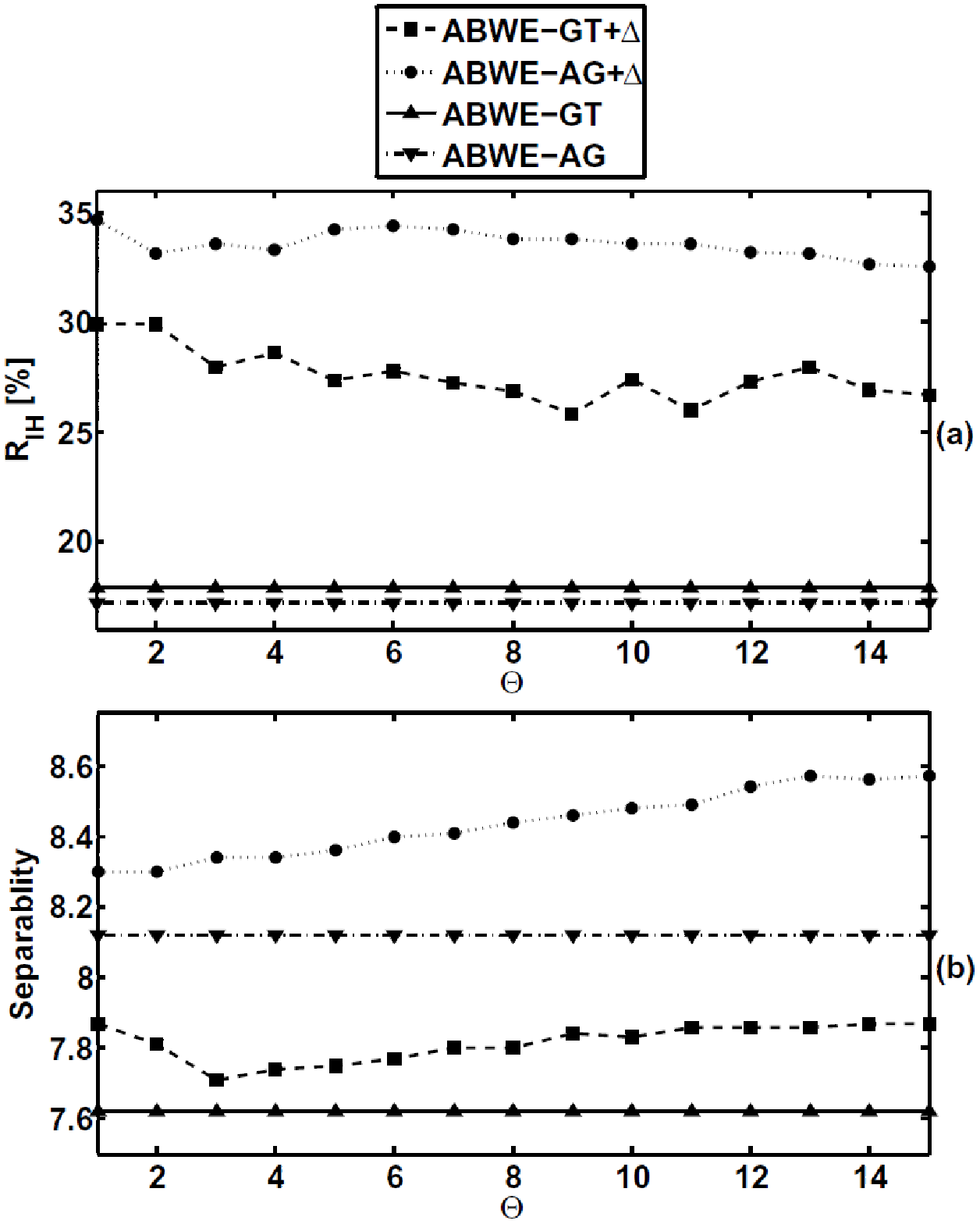

Figure 4.1: *Plot of ratios ($\widehat{R}_{IH}$) for global (ABWE-GT) and age-specific (ABWE-AG) models with and without Δ (a). Plot of Separability ($\varepsilon(\mathbf{x})$) for global (ABWE-GT) and age-specific (ABWE-AG) models with and without Δ (b). Half window size, Θ is selected between range of 1 to 15 to compute Δ.*

As demonstrated in the earlier tables and figures, the use of class-specific, age-specific and delta information indeed improves the ratio and separability values. Therefore these can be exploited for developing ABWE methods. The development of different ABWE methods using these information is explained in the following sections.

4.2 ABWE using Class-Specific Information

In the previous chapter, a significant improvement in the recognition performance of narrowband children's speech in mismatched condition (i.e, against adults acoustic models) was noted with ABWE. In that work narrowband speech is artificially extended by GMM based ABWE algorithm which exploits the correlation between NB and HB speech signals. The mutual information between NB and HB speech varies widely across different speech classes. Thus the use of class specific ABWE transformation is expected to result in further improvement. Motivated by this, in this study we explore methods for applying class-specific ABWE transformations in the context of mismatched children's speech recognition. In contrast to a similar nature work [105], this work explores finer (non-phonetic) unsupervised classification of speech and children's ASR context.

4.2.1 Derivation of Class Information for ABWE Transformation

It is well known that mutual information between the narrowband and the higher band spectra varies across the speech signal. So the use of global ABWE transformer may not be optimal. It would be better to train different ABWE transformer for each of the classes present in speech. Given a speech signal, the derivation of the class information is a non-trivial problem. The phonetic/word labelling of the speech signal is the natural way to incorporate the speech class information for ABWE, but for deriving the phone/word labels, one is required to do the phone/word recognition prior to ABWE of the narrowband signals. It not only increases the complexity of the overall ABWE system, but also limited by the levels of accuracy achieved by the phone/word recognizer. On the other hand there are some methods reported in literature which employ data driven clustering for ABWE instead of using the phone/word classification.

In the following, we describe two approaches explored for the incorporation of class information for ABWE. The first one is an unsupervised non-phonetic data driven classification approach and the other one is supervised approach using the oracle word level classification. The purpose of the later approach is to benchmark the performance of class specific ABWE transformation.

4.2.1.1 Unsupervised Classification

For the unsupervised classification, we have used an algorithm proposed in [122]. It uses an HMM based unsupervised joint feature analysis framework to build a correlation model between narrowband and higher band speech. The HMM based unsupervised classifier is used to jointly segment temporal and spectral features. The joint temporal and spectral feature patterns are used to form a correlation model between NB and HB speech.

An HMM ($\Gamma^{\mathbf{Z}}$) is trained using the joint feature $\mathbf{Z}$, to capture recurrent phonetic segments in speech. The HMM ($\Gamma^{\mathbf{Z}}$) is composed of N parallel HMMs, $\{\gamma_1^{\mathbf{Z}}, \gamma_2^{\mathbf{Z}}, \ldots, \gamma_N^{\mathbf{Z}}\}$, where $\gamma_N^{\mathbf{Z}}$ is chosen to be a single state HMM $\{S_N\}$ with diagonal covariance GMM. Given the joint feature sequence, $\{\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_M\}$ where $\mathbf{Z}_M$ denotes the joint feature vector for M^{th} frame, the segmentation of the feature sequence is performed using the Viterbi decoding to maximize the probability of model match. The Viterbi decoding yields a state sequence $\{S_1^{\mathbf{Z}}, S_2^{\mathbf{Z}}, \ldots, S_M^{\mathbf{Z}}\}$ associated with the feature sequence $\mathbf{Z}$. The correspondence between states $S_M^{\mathbf{Z}}$ and feature vectors $\mathbf{Z}_M$ forms vector quantization like classification which also benefits from temporal correlation due to the nature of the HMM structure. For class-specific ABWE transformation, the HMM based data driven unsupervised classification is performed for sufficient large amount of development speech data and then separate GMM based ABWE transformer is trained for each of the data driven classes.

For deriving the class information for the narrowband speech signal to be bandwidth extended the following procedure is used:

(i) The NB features $\mathbf{X}$ are extracted from the given narrowband speech.

(ii) As the joint feature based HMM model $\Gamma^{\mathbf{Z}}$ uses GMM with diagonal covariance matrices, it allows the extraction of the HMM model $\Gamma^{\mathbf{X}}$ for the narrowband feature by simple

partition of joint feature vector.

(iii) Temporal segmentation of the NB feature sequence $\mathbf{X}$ is derived by decoding with the HMM model $\Gamma^{\mathbf{X}}$ to extract the sub-phone patterns with a state sequence $\mathbf{S}^{\mathbf{X}}$.

(iv) Given the sub-phone classes (states) the class specific GMM based ABWE transformers are trained.

4.2.1.2 Supervised Classification

In supervised classification, the true word level information is used for both learning the class-specific ABWE transformation and for the narrowband speech signal to be bandwidth extended. Thus this approach cannot be used in general where the true transcription may not be available. But it does provide us a useful benchmark to the performance improvement possible with the use of class specific ABWE transformations. The procedure for incorporation of the class information in learning of ABWE transformers and during bandwidth extension of NB signal are as follows:

(i) The training set narrowband speech data is forced aligned with its true word level transcriptions using an existing narrowband speech trained recognition system.

(ii) Separate class specific GMM based ABWE transformer are then developed using both NB and HB data.

(iii) For the test data, that is, NB speech to be bandwidth extended, the NB features are computed and force aligned with the true transcriptions on the existing NB speech recognition system to derived the class information.

(iv) Based on the class information derived, the appropriate class specific GMM based ABWE transformer is used to estimate the HB features.

4.2.2 ASR using Class-Specific Information based ABWE

In this work we have developed three different kinds of ABWE systems: global transformation based ABWE (ABWE-G, same as described in previous chapter for ABWE), unsupervised

class specific transformation based ABWE (ABWE-UNSUP-CLS), and supervised class specific transformation based ABWE (ABWE-SUP-CLS). In ABWE-G systems, a single GMM based ABWE transformer is used. The different parameters chosen are: window length of 20 ms, window shift of 10 ms, the narrowband LPC order of 10, the wideband LPC order of 20. In ABWE-UNSUP-CLS system, the speech signal is segmented into 128 classes through an HMM model. In ABWE-SUP-CLS system, the speech signal is segmented into 12 classes (zero to nine digits, oh, /sil/) by forced alignment with true word level transcriptions. In all three systems, the ABWE transformation(s) are trained. Also the data used for training the ABWE transformer for children's narrowband speech was kept mutually exclusive to the children's test sets. The connected digit recognizer is developed using the procedure described in Section 3.1 of Chapter 3.

4.2.2.1 Results and Discussion

As the number of classes in different systems are widely different to the data available for training of ABWE transformer would also differ significantly if same size GMM models are used in ABWE systems. Therefore, we have first explored the appropriate size of full covariance GMM models for all ABWE systems. The different speech quality measures used for the performance evaluation are described in Appendix A.

The performances of different ABWE systems developed for varying size of GMM for children's test set are given in Table 4.6. For the assessment of the performance of ABWE systems in detail, different speech quality measures are also computed along with the ASR performance of bandwidth extended children's speech under mismatch condition. With the tuning for appropriate GMM size, we note that ABWE-UNSUP-CLS and ABWE-SUP-CLS systems have resulted in best WER of 4.88% and 4.78%, respectively for single Gaussian, while ABWE-G has resulted in WER of 6.23% with 16 Gaussian mixture. Further we note that the performance in terms of different speech quality measures also correlate with that of ASR performance with ABWE.

Table 4.7 summarizes the relative improvement in narrowband children's speech recognition

performance with different ABWE systems. It is to note that ABWE-UNSUP-CLS and ABWE-SUP-CLS systems provides a relative improvement of 21.67% and 23.27% over that of ABWE-G system. Apart from the WER, several objective measures are also computed for each of the studies. The objective measures closely match with the WER, when class-specific information is used for modelling information between NB and HB cases. This shows the significance of using class-specific information compared to the existing global transformation method.

Table 4.6: *Performances of different ABWE systems developed for varying size of GMM in ABWE systems for children's test set . The performances are measured in terms of %WER under mismatched condition as well as different speech quality measures such as log likelihood ratio (d_{LLR}), weighted slope metric (d_{WSM}), likelihood ratio (d_{LR}), cepstrum distance (d_{CEP}), weighted likelihood ratio (d_{WLR}), root mean squared log spectral distortion (d_{LSD}), segmental signal to noise ratio in dB (segSNR).*

No. of Gaussians	ABWE-G							
	WER%	d_{LLR}	d_{WSM}	d_{LR}	d_{CEP}	d_{WLR}	d_{LSD}	segSNR
1	6.64	0.4008	2.51	0.7238	**0.1652**	0.07363	0.6598	-3.09
2	6.34	0.4015	**2.50**	0.7108	0.1681	0.07176	0.6568	-3.08
4	6.44	0.3997	2.55	**0.7021**	0.1689	**0.07105**	0.6650	-3.08
8	6.76	**0.3971**	2.59	0.7416	0.1669	0.07117	0.6598	-3.00
16	**6.23**	0.4004	2.53	0.7275	0.1671	0.06983	**0.6419**	-3.09
32	8.61	0.4054	2.59	0.7225	0.1623	0.07412	0.7390	-3.01
No. of Gaussians	ABWE-SUP-CLS							
	WER%	d_{LLR}	d_{WSM}	d_{LR}	d_{CEP}	d_{WLR}	d_{LSD}	segSNR
1	**4.78**	**0.3470**	2.47	**0.5581**	**0.1478**	**0.05543**	**0.6049**	-3.09
2	4.85	0.3539	**2.41**	0.5813	0.1522	0.05585	0.6082	-3.10
4	4.83	0.3529	2.49	0.5742	0.1516	0.05455	0.6100	-3.11
8	4.81	0.3510	2.50	0.5928	0.1517	0.05748	0.6189	-3.09
16	5.59	0.3647	2.54	0.5820	0.1514	0.05853	0.6451	-3.04
No. of Gaussians	ABWE-UNSUP-CLS							
	WER%	d_{LLR}	d_{WSM}	d_{LR}	d_{CEP}	d_{WLR}	d_{LSD}	segSNR
1	**4.88**	**0.3579**	**2.49**	**0.5821**	**0.1527**	**0.06162**	**0.5970**	-3.06
2	5.18	0.3613	2.51	0.5845	0.1539	0.06252	0.6055	-3.11
3	5.09	0.3648	2.52	0.5956	0.1552	0.06340	0.6116	-3.13
4	5.21	0.3709	2.53	0.6147	0.1570	0.06523	0.6158	-3.11
6	5.37	0.3750	2.54	0.6346	0.1596	0.06775	0.6180	-3.16
9	5.86	0.3757	2.56	0.6373	0.1579	0.06458	0.6215	-3.08
12	6.09	0.3789	2.57	0.6461	0.1586	0.06525	0.6203	-3.08

Table 4.7: *Recognition performances for children's test set with narrowband (NB), wideband (WB) and ABWE transformed test data conditions. The ABWE transformed data conditions include use of global (G); unsupervised class specific (UNSUP-CLS) and supervised class specific (SUP-CLS) transformations.*

Children's test set condition	Adults' acoustic models condition	WER%
NB	NB	9.37
WB	WB	3.21
ABWE-G	WB	6.23
ABWE-UNSUP-CLS	WB	4.88
ABWE-SUP-CLS	WB	4.78

4.3 Feature Domain MFCC based ABWE

The traditional ABWE approaches were mostly based on LPC features or its variants such as line spectral frequency (LSF) for modelling the relation between LB and HB portions of the speech spectra. Despite the successful use of MFCC features in many other speech signal processing tasks, it was not considered for ABWE purposes till lately. The main reason for that neglect was the process of conversion from MFCC domain to speech domain being rather tedious than that is from LPC domain. The first work exploring the MFCC representation of speech in the ABWE modelling was reported recently [139]. In subsequent works [90], it was shown that the MFCC features result in enhanced mutual information compared to that with LSF/LPC features. In [117], a study on the use of different types of features for NB in ABWE also showed that MFCC features give the best separability while having a high mutual information between lower and higher bands of the spectra. In all of these works, the ABWE performance with MFCC feature were evaluated in context of adults' speech only.

Motivated by large differences between adults and children's speech, in this work we explore the MFCC based ABWE approach for the bandwidth extension of children's speech. The main purpose is to augment the existing adaptation/normalization techniques in bridging the large gap between the ASR performance for NB and WB speech. We have also presented a novel method for deriving the bandwidth extended (BWE) MFCC features directly exploiting the MFCC based ABWE framework. The proposed method has much lower computational com-

plexity as it avoids the need for conversion to speech domain for parameterization of extended speech for ASR purpose without any loss in the performance. In another work [140], we have experimented with creation of class-specific ABWE using unsupervised clustering. With the use of unsupervised class-specific ABWE, not only the mutual information between LB and HB get enhanced, but also it results in improved performance for ASR of the children's speech. Motivated by that we have also studied the age-specific conditioning of ABWE and the effect of inclusion of delta features in ABWE of children's speech.

4.3.1 Novel Feature domain ABWE modeling

In this section, we describe the procedure followed for developing the MFCC features based ABWE framework following the one originally proposed in [139]. This technique basically exploits the compact representation produced by the mel filter-bank for defining the LB and the HB portion of speech spectra. For developing ABWE model, the wideband speech is analyzed to produce the short-time magnitude spectra which are then multiplied with 22-channel mel filter-bank and the energies at the output of the filters are log-compressed. Out of these 22 energies, the ones from channels 1-15 are grouped to represent the LB spectra and the remaining channels excluding 16 are grouped to represent the HB spectra. The channel 16 is dropped as it falls both in LB and HB ranges. These grouped energies are then converted into 15-dimensional MFCC features (C_0-C_{14}) for LB and 6-dimensional MFCC features (C_0-C_5) for HB by taking the discrete cosine transform (DCT). Finally for representation purpose, the LB features are truncated to 10-dimensional MFCC features (C_0-C_9) while HB features are kept as it is. For enhancement of spectra, 16-dimensional MFCC feature obtained by joining the MFCC features corresponding to LB and HB are modelled using GMM. For adjusting the energy of the estimated HB spectra in reference to given LB spectra as suggested in [139], the ratio between mean energies of HB and LB spectra is computed and a separate GMM is trained on 11-dimensional joint feature consisting of gain (G) and LB MFCC features.

Given the LB MFCC features, the HB MFCC features are determined using the minimum mean square error (MMSE) estimation criterion as originally proposed in [12] and also detailed

in our previous work [141]. The similar procedure is used for estimation of the gain factor. The detailed block diagram of obtaining the LB and WB features from a wideband development data for the ABWE modelling is shown in Figure 4.2(a).

4.3.2 Efficient derivation of extended wideband MFCC

As stated earlier in this work, the motivation for the development of the ABWE of narrowband children's speech is not to enhance the quality of its perception but to improve the ASR performance for the narrowband children's speech. So after obtaining the bandwidth extended speech from the given narrowband speech, it is converted into MFCC features following the standard procedure for the ASR purpose. With the use of MFCC based ABWE scheme, the derivation of the wideband MFCC features get significantly facilitated. The complete flow graph of the bandwidth extended wideband MFCC computation is illustrated in Figure 4.2(b).

To begin with, the given 8 kHz sampled narrowband signal is interpolated by a factor of 2. The interpolated narrowband signal is analyzed to produce the short-time magnitude spectra which are multiplied with the same 22-channel mel filterbank as used in generating the MFCC features for the ABWE modelling. As in the interpolated narrowband signal the spectral content be limited to lower band region. Thus the log-compressed energies from channels 1-15 are considered only. These 15-dimensional log-energies are then converted using DCT into 10-dimensional MFCC features (C_0-C_9) to match with the LB representation used in ABWE modelling. Given the LB MFCC features, the GMM based map estimates 6-dimensional HB MFCC features which are converted to 6-dimensional HB log-energies by taking inverse DCT. After obtaining the HB log-spectra vector ($\mathbf{E}_{HB}$), its energy is adjusted version $\tilde{\mathbf{E}}_{HB}$ relative to the observed LB log-spectra vector ($\mathbf{E}_{LB}$) obtained by using the gain factor (G) estimated separately corresponding to the observed LB MFCC features as

$$\tilde{\mathbf{E}}_{HB} = \left(\frac{G \times \bar{E}_{LB}}{\bar{E}_{HB}} \right) \mathbf{E}_{HB} \tag{4.10}$$

where $\bar{E}_{HB}$ and $\bar{E}_{LB}$ denote the mean value of HB and LB spectral vectors, respectively.

For the generation of the extended full band mel spectra, prior to concatenating 15-dimensional

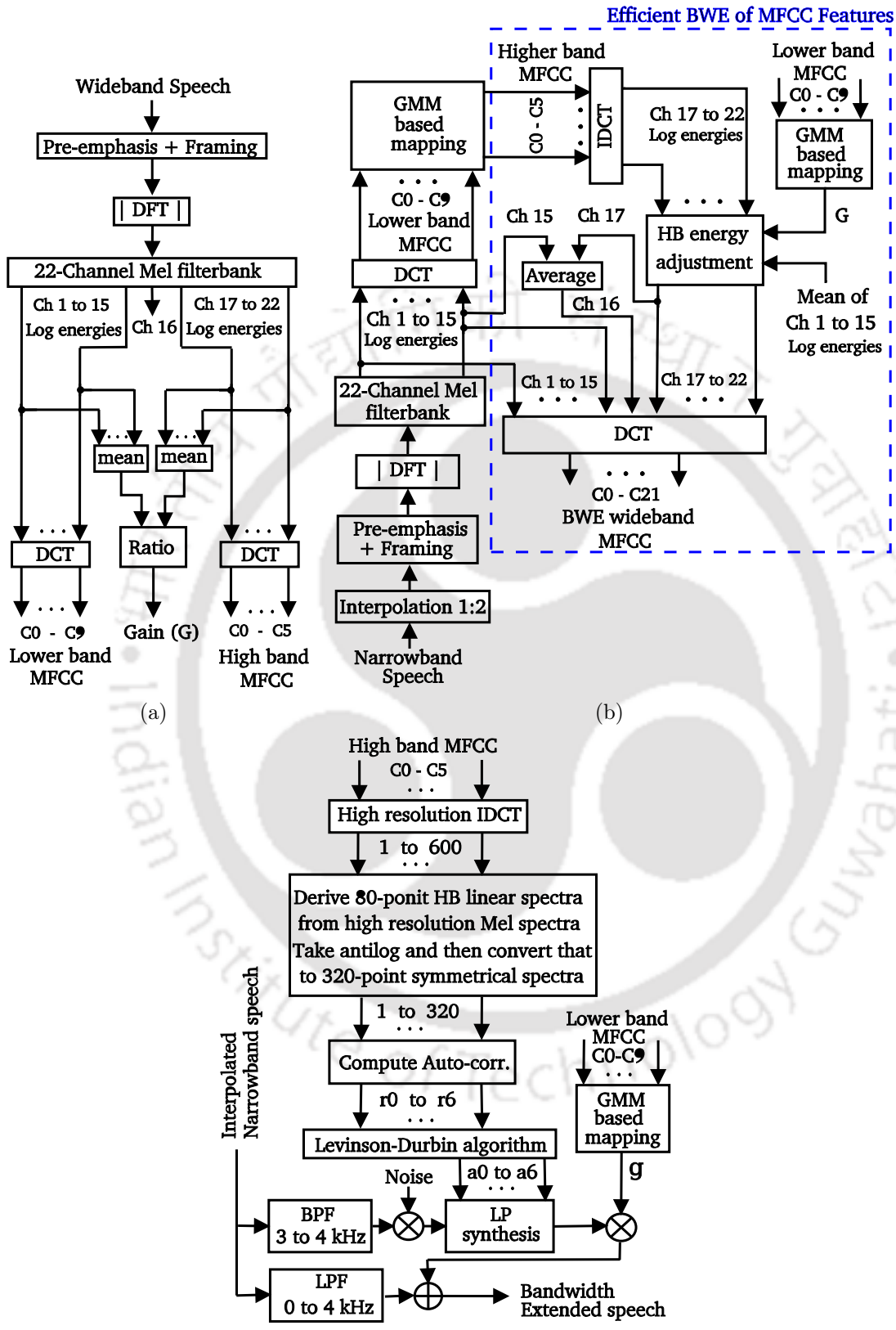

Figure 4.2: The detailed block diagrams of the proposed and the default (speech domain) MFCC based ABWE approaches. (a) Derivation of LB and HB MFCC features and the gain factor between LB and HB for ABWE modelling. (b) Proposed approach of direct computation of bandwidth extended MFCC features for ASR purpose exploiting the MFCC based ABWE approach. (c) Additional processing involved in default speech domain MFCC based ABWE approach used for contrast purpose.

LB log-energies corresponding to the given interpolated narrowband signal with energy adjusted 6-dimensional HB log-energies, the log-energy corresponding to the dropped 16^{th} channel is re-estimated by averaging the log-energies of channels 15 and 17. The extended full band mel spectra is then converted into 22-dimensional MFCC features using DCT.

For contrasting the performance of the proposed approach with the default way, we have also derived the bandwidth enhanced speech following the procedure as described in [139]. Figure 4.2(c) shows the block diagram of the procedure required for the derivation of the bandwidth enhanced wideband speech given the enhanced MFCC features. After conversion to the speech domain, additional processing is required to compute the MFCC features for ASR purpose. Thus, on comparing Figs. 4.2(b) and 4.2(c), it can be easily assessed that a huge reduction of computational complexity is obtained with proposed approach in context of ASR. It is also worth highlighting that the LP modelling involved in existing speech domain approach may result in degraded final MFCC features in particular for the unvoiced speech.

4.4 Delta Features and Age Information for MFCC based ABWE

In this section we describe about two types of enhancements studied in ABWE of children's speech. The one is the inclusion of delta features and the other is the age-specific conditioning in ABWE modelling. To access the impact of above two enhancements for children's speech case, the normalized mutual information (MI) [142] between LB and HB is also computed.

4.4.1 Inclusion of Delta Features in ABWE

The recent studies [89, 90, 143] have shown that the inclusion of delta features in MFCC features for ABWE modelling helps to improve the MI between LB and HB. This motivated us to explore the same in the context of ABWE of children's speech. As per HTK implementation, Θ specifies the half length of window used in the computation of delta features. The formula for the delta feature is

$$\delta_t = \frac{\sum_{l=1}^{\Theta} l.(c_{t+l} - c_{t-l})}{2\sum_{l=1}^{l=\Theta} l^2} \tag{4.11}$$

Table 4.8: *Age-wise break up of children's digit data in the development and the test sets.*

Age (yrs)	6	7	8	9	10	11	12	13-14	15
Dev	253	752	3332	1893	1676	4995	1065	506	253
Test	-	1012	711	1902	2116	1823	2477	506	253

where δ_t is a delta coefficient at frame t, c_{t+l} is the corresponding static features at frame $t+l$, and Θ specifies the half length of window used. For this study, the included memory is controlled by varying the half window length Θ from 1 to 15 in steps of 2. Let $\mathbf{X}$ and $\mathbf{Y}$ represents LB and HB static MFCC feature vectors and their corresponding delta feature vectors are represented by $\Delta\mathbf{X}$ and $\Delta\mathbf{Y}$ respectively, then the joint MFCC features including deltas as used for ABWE modelling is given by $\mathbf{Z} = \left[\mathbf{X}^T, \Delta\mathbf{X}^T, \mathbf{Y}^T, \Delta\mathbf{Y}^T\right]^T$.

4.4.2 Age-specific conditioning in ABWE

As a result of large developmental changes in children, the children's speech is characterized to have higher variability unlike that in adults' speech. The values of acoustic attributes like pitch, formant frequencies, and speaking rate exhibit significant changes across children's of differing age groups [40]. To account for this fact, we analyzed the impact of conditioning the speech data used for training of GMM based ABWE model. For this study, a separate ABWE model is developed for each age-group of the speakers in the training set of TI-DIGITS corpus. The age-wise break up of the total number of children speakers in the training and test sets in Table 4.8.

Distribution of data across the children's age groups is as fallows: 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 years age groups have train set of 253, 752, 3332, 1893, 1676, 4995, 1065, 0, 253 digits, respectively and test set of 00, 1012, 711, 1902, 2116, 1823, 2477, 253, 253, 253 digits, respectively. The same is summarized in Table 4.8.

4.4.2.1 ABWE models

In this work we have created four different kinds of ABWE models: global transformation based ABWE (ABWE-GT), global transformation based ABWE including delta features (ABWE-GT+Δ), age-specific transformation based ABWE (ABWE-AG), and age-specific trans-

Table 4.9: *Performances for children's test set for the default (speech domain) and the proposed (feature domain) approaches using ABWE-GT.*

ABWE-GT		WER (%)
Default	with gain	9.19
Proposed	without gain	9.00
	with gain	8.88

formation based ABWE including delta features (ABWE-AG+Δ). For training the different types ABWE models, the children's data from TI-DIGITS is used which is mutually exclusive to the children's test set.

4.4.2.2 ASR system

For the adults' speech trained digit recognition system, the WER for adults' test set turned out to be 1.28% and 0.35% for NB and WB speech cases, respectively. Table 4.9 shows the performances of the proposed MFCC based ABWE approach for with and without gain adjustment along with existing MFCC ABWE approach in the context of children's ASR. On comparing with the existing MFCC approach, the proposed approach does not found to give any degradation in performance at the same time is quite efficient. From the computational efficiency point of view the system without gain adjustment is used.

The performances of bandwidth enhancement of children's test set using ABWE-GT and ABWE-AG approaches when tested on WB adults' speech trained ASR models are given in bottom row of Table 4.10 along with that of original WB and NB children's speech cases. On comparing we note that global and the age-specific approaches have resulted in 0.83% and 2.5% absolute improvement respectively over NB baseline. The age-specific conditioning of ABWE results in 1.7% absolute improvement in performance over that of global ABWE. Note the performance for ABWE-AG is evaluated in supervised mode i.e., using the true age of test children speakers. This result only provides an assessment of the sensitivity of ABWE modelling to the high variability in children's speech. For actual use of this fact, it is required to estimate of the ages of the test children speaker in unsupervised manner and that would be addressed in future study.

Table 4.10: *Performances for using ABWE-GT and ABWE-AG on children's test along with age-wise breakup.*

WER %				
Age in yrs	ABWE-GT	ABWE-AG	NB	WB
07	18.68	13.93	22.53	6.92
08	19.55	18.14	16.46	5.63
09	16.77	12.67	18.98	6.20
10	4.96	6.00	5.29	0.80
11	5.70	2.69	4.83	2.25
12	4.28	4.00	5.73	1.49
13-14	0.99	0.40	1.78	0.00
15	1.98	1.58	1.98	1.19
Avg.	9.00	7.33	9.83	3.02

Table 4.11: *Performances for varying delta MFCC features used in global (ABWE-GT+Δ) and age-specific (ABWE-AG+Δ) models.*

Condition	WER (%) for the inclusion of varying memory Δ														
Half window size, Θ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ABWE-GT+Δ	10.04	9.52	10.09	10.06	9.18	8.40	9.34	9.83	6.23	**5.41**	8.80	9.03	8.77	9.25	8.93
ABWE-AG+Δ	8.75	8.55	7.93	8.03	6.93	6.78	6.11	5.78	5.69	5.42	**5.15**	5.51	5.20	5.19	5.29

The performance of ASR system using delta features information is given in Table 4.12. As it can be observed, the delta features significantly improves the performance in both cases of global transform and age-specific cases.

The details of different WER obtained for different values of Θ for computation of Δ features are given in the Table 4.12. As it can be observed the good performance is achieved in case of lower age children for larger values of Θ. Alternatively, good performance (lower WER) is achieved for higher age children for smaller values of Θ.

Table 4.12: *Performances of ABWE-GT+Δ (global transform) and ABWE-AG+Δ (age specific transforms) systems for children's test set partitioned by age. Half window size, Θ is selected between range of 1 to 15 to compute Δ. The performances are measured in terms of WER% under mismatched condition.*

ABWE-GT+Δ															
WER%															
Age in years	Half window size, Θ														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
7	20.06	19.76	20.36	20.26	18.97	17.98	19.37	20.06	14.43	**12.75**	17.79	18.28	18.08	18.28	18.08
8	19.83	18.71	20.11	20.68	18.99	17.44	19.41	21.11	13.50	**10.41**	19.13	19.27	19.13	19.41	18.42
9	18.46	17.51	19.09	18.19	16.98	15.72	17.40	18.35	12.20	**11.20**	16.67	17.03	16.56	17.35	16.46
10	5.81	5.58	5.72	5.91	4.87	4.25	4.82	5.15	2.98	**2.36**	4.68	4.91	4.73	5.07	4.96
11	6.86	6.20	6.64	6.97	6.36	5.70	6.36	6.64	3.57	**3.29**	5.92	6.03	5.38	5.87	5.98
12	4.97	4.76	4.88	5.01	4.32	3.79	4.52	4.88	2.62	**2.10**	3.92	4.12	4.24	4.72	4.36
13	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.00	**0.00**	0.00	0.00	0.00	0.00	0.00
14	3.95	3.16	3.56	3.16	3.56	3.16	3.56	3.95	0.79	**0.79**	3.16	3.16	1.98	3.56	3.56
15	1.98	1.58	1.98	1.58	1.98	1.98	1.58	1.98	1.58	**1.58**	1.98	1.98	1.98	1.98	1.98
Avg.	10.04	9.52	10.09	10.06	9.18	8.40	9.34	9.83	6.23	**5.41**	8.80	9.03	8.77	9.25	8.93

ABWE-AG+Δ															
WER%															
Age in years	Half window size, Θ														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
7	21.15	21.54	17.09	19.17	14.82	12.75	13.44	12.94	**11.96**	12.35	12.25	12.55	12.55	12.55	12.35
8	21.52	18.71	16.74	14.35	15.33	14.91	10.27	15.33	17.72	12.10	12.38	11.53	11.81	10.27	**9.28**
9	14.88	12.93	14.35	18.30	13.88	13.62	13.46	10.78	10.78	10.99	**10.09**	11.83	11.04	11.09	11.25
10	4.25	4.16	5.39	3.21	3.97	3.36	3.17	2.88	2.50	2.22	1.89	1.65	**1.47**	1.56	2.27
11	5.60	5.76	4.44	3.89	3.62	4.77	3.02	3.07	**2.47**	2.80	3.02	3.29	2.63	2.63	2.52
12	3.96	5.13	3.63	3.15	2.79	3.03	2.75	2.26	2.34	2.50	**2.10**	2.46	2.30	2.54	2.70
13	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
14	**0.79**	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79
15	**1.19**	1.58	1.58	1.58	1.58	1.19	1.19	1.58	1.58	1.19	1.19	1.19	1.19	1.19	1.19
Avg.	8.75	8.55	7.93	8.03	6.93	6.78	6.11	5.78	5.69	5.42	**5.15**	5.51	5.20	5.19	5.29

4.4.3 Estimation of Age-specific information

The approach we propose here is inspired by the work of Pellom and Hansen [144], whose algorithm relies on a set of height-dependent GMMs modelling MFCC distributions based on the fact that MFCCs correlate with speaker height [145].

Each age group training data lower band features ($\mathbf{X}$) is modeled into a GMM represented as λ. The PDF of $\mathbf{X}$ is modeled as a mixture of M n-variate Gaussian PDFs,

$$P\left(\mathbf{X}|\lambda\right) = \sum_{i=1}^{M} \alpha_i b_i(\mathbf{X}), \quad \sum_{i=1}^{M} \alpha_i = 1 \text{ and } \alpha_i \geq 0 \tag{4.12}$$

where $b_i(\mathbf{X})$ and $\alpha_i, i = 1, ..., M$ are the component densities and the component weights, respectively. Each component density is a n-variate Gaussian function of the form:

$$b_i(\mathbf{X}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{C}_{xx_i}|^{\frac{1}{2}}} e^{-\frac{1}{2}\left[\left(\mathbf{X}-\boldsymbol{\mu}_{x_i}\right)^T \mathbf{C}_{xx_i}^{-1}\left(\mathbf{X}-\boldsymbol{\mu}_{x_i}\right)\right]} \tag{4.13}$$

with $\boldsymbol{\mu}_i$ is $n \times 1$ mean vector and $\mathbf{C}_{xx_i}$ is $n \times n$ covariance matrix with

$$\widehat{\mathcal{A}} = \arg\max_{\mathcal{A}} P\left(\mathbf{X}|\lambda_{\mathcal{A}}\right) \tag{4.14}$$

where $\mathcal{A}$ is the age of the children speakers in years. The age-wise break up of children's development data is shown in Table 4.8. Each age group of development data is used to train a separate GMM $\lambda_{\mathcal{A}}$. $\widehat{\mathcal{A}}$ estimated age computed from a speech vector $\mathbf{X}$ and age specific GMMs $\lambda_{\mathcal{A}}$.

Each frame's probable age group is determined using maximum likelihood (ML). Then age group of a utterance is found by using *mode* of frames belong to speech regions.

The performance of ASR system using the estimated age-specific information is tabulated in Table 4.13. The performance is comparable to the case of supervised age-specific information. This infers that, the GMM for age prediction have indeed modelled the age information well and hence resulted in the good performance.

Table 4.13: *Performances for using ABWE-GT, ABWE-AG and ABWE-AG-ML on children's test along with age-wise breakup.*

WER %					
Age in yrs	ABWE-GT	ABWE-AG	ABWE-AG-ML	NB	WB
07	18.68	13.93	14.33	22.53	6.92
08	19.55	18.14	16.46	16.46	5.63
09	16.77	12.67	13.77	18.98	6.20
10	4.96	6.00	4.40	5.29	0.80
11	5.70	2.69	4.61	4.83	2.25
12	4.28	4.00	3.35	5.73	1.49
13-14	0.99	0.40	0.40	1.78	0.00
15	1.98	1.58	1.58	1.98	1.19
Avg.	9.00	7.33	7.31	9.83	3.02

4.5 Summary

In this chapter new methods for ABWE using auxiliary information were proposed. The auxiliary information including class (digit) specific, age specific and delta features related information. To motivate the work initially, the statistical measures are computed. These include mutual information, differential entropy, their ratio and separability measure. The important observation that motivated the works proposed in this chapter is the variability in the statistical measure values across class, age and delta features. Also, it was observed that the statistical measure values increase when the auxiliary information is exploited as compared to the global transform case. Based on this observation it was motivated to exploit the auxiliary information for ABWE. artificial bandwidth extension The first method of ABWE was based on exploiting the class (digit) specific information. To see the effectiveness of the class specific information, initially supervised class specific ABWE method is proposed where the class information is taken from the available transcription. The performance (WER) for NB is 9.37%, WB is 3.21%. Therefore the best performance we can expect after ABWE is the result of WB case. The global transform case gives 6.23% and the supervised case gives 4.78%. After an unsupervised class-specific ABWE method is developed where the class information is predicted from the test data and used for ABWE and this provided 4.88%. The results are along the expected lines inferring that the class specific information is indeed effective in performing ABWE.

A feature domain ABWE method is proposed using MFCC representation. The existing ABWE method based on MFCC follows the vocoder framework where the speech is synthesized from MFCC and then the ABWE is performed in vocoder framework. Alternatively, since the goal of present work is ASR, the ABWE using MFCC can be significantly simplified by working in the feature domain itself. Based on this motivation, a feature domain ABWE method using MFCC is proposed. The existing ABWE method based on MFCC provides a WER of 9.19% where as the proposed one provides a WER of 8.88%. This shows that even though improvement in terms of performance it is moderate, but the same is achieved with much higher simplified

methodology.

The proposed feature domain MFCC based ABWE methods is used for further demonstrating the significance of age specific and delta information. The lower bound and upper bound in performance for age specific study is limited by the NB and WB cases giving 9.83% and 3.02%. The global transform case gives a WER of 9.0%. The supervised case of age specific information provides a WER of 7.33% which is the best performance that can be achieved using age information. The unsupervised age prediction method provides a WER of 7.31% which is comparable to that of supervised case.

The next study exploited the delta features information on top of the age specific information. The delta features information improves the performance of the system from 9.0% to 5.41% for global transform case and from 7.33% to 5.15% in case of age specific information case.

Thus all the above studies infer that it is indeed possible to develop ABWE methods using the auxiliary information and provide significantly better performance compared to the NB case.

Having exploited the auxiliary information for ABWE, the next chapter focuses on developing ABWE using the recent developments in the signal processing field. The sparse representation based signal modelling approach is found to result in state-of-the-art performances in many signal processing applications [132, 133]. Motivated by this, the present work also explores developing an ABWE method exploiting the sparse representation modelling of speech frames in the signal domain. The main merit of this approach is it is a non-parametric ABWE approach and also does not make use of the mostly followed source-filter model approach for ABWE.

5

Proposed Sparse Representation based ABWE

Contents

This chapter presents a novel approach to ABWE based on sparse representation of speech signals. The proposed approach is motivated by the premise that a redundant dictionary whose atoms are more speech-like can produce a sparse representation of speech signals. As a consequence of that such a dictionary is expected to be somewhat consistent in sparse coding of the narrowband and the wideband speech signals. This attribute can be exploited for the bandwidth extension purpose. In this work, the dictionaries for sparse representation are created using K singular value decomposition (KSVD) algorithm. The proposed approach is found to be less effective for unvoiced speech and the efforts made to address the same are also described. On comparing with the existing line spectral frequency based ABWE method, the proposed ABWE approach is found to give better performance in terms of speech quality measures. In the later part, we are exploring an existing semi-coupled dictionary learning (SCDL) algorithm for ABWE (SC-ABWE), which was proposed for image-style transformation application. SCDL algorithm learns bidirectional transformation iteratively with dictionary learning. These dictionaries are not fully coupled and hence provide more freedom for the transformation. We found an improvement in the SC-ABWE performance in terms of objective quality measures.

The traditional ABWE approaches are based the source-filter model of speech. In addition to that, for modelling the relation between LB and HB portions of the speech, all traditional approaches employ a mapping function which is based on either vector quantizer (VQ) [11] or Gaussian mixture model (GMM) [12] or artificial neural network (ANN) [14]. The mapping functions are developed based on some kind of features of speech either as linear predictive coefficient (LPC) [12] or its variant line spectral frequency (LSF) [146] or mel frequency cepstral coefficient (MFCC) [139].

In recent past, the sparse representation (SR) based signal modelling approach has received a lot of attention. In a number of signal processing domains, the SR approach is found to result in the state-of-the-art performances [132]. Motivated by that in this work we present an initial study exploring the sparse representation for ABWE purpose. The proposed ABWE approach is developed by exploiting the sparse representation modelling of speech frames in the signal domain. Thus it is a non-parametric ABWE approach and also does not make use

of the ubiquitous source-filter model.

In image processing literature, we came across a proposal for developing coupled dictionaries for sparse coding in cross-style image synthesis domains. The examples for cross-style synthesis domains include converting a low resolution image to a high resolution one, matching face sketch of person to her photo and so on. In [133], the signals/ features corresponding to the cross-style domains are joined into a vector and a single dictionary is created by simultaneously minimizing the representation error in both domains under the constraint on sparsity. On stream-wise splitting of so trained dictionary results in two dictionaries having a coupling among their atoms. In subsequent works, referred to as *semi-coupled dictionary*, a linear transformation between the sparse codes is also learned along with the dictionaries to address the mismatch in the sparse coding for cross-style synthesis domains.

Sparse and redundant data modelling seeks the representation of signals as linear combination of a small number of atoms from a data driven dictionary. Recently, there is a fast increasing interest in dictionary application. Dictionary learning methods mainly focus on training an over-complete dictionary in a single feature space for various recovery or recognition tasks. In many applications and scenarios, we have coupled sparse feature spaces. Coupled sparse dictionary based methods for image processing applications like, image style transformation and super resolution, have reported state-of-the-art performances. Coupled sparse dictionary methods are based on the motivation that the two fields of the data can be represented by the same set of sparse codes. The method essentially concatenates the two feature spaces and converts the problem to the standard sparse coding in a single feature space. As such, the resulting dictionaries are not indeed trained for each of the feature spaces individually. Sparse codes of one feature sparse can therefore synthesize the other feature space with the help of coupled dictionaries.

To overcome the limitation of coupled dictionaries, semi-coupled dictionaries (SCDL) have learnt simultaneously the dictionary pair and a mapping function. The pair of dictionaries aims to characterize the two structural domains of the two fields, and the mapping function is to reveal the intrinsic relationship between the two styles for synthesis. In the learning process, the

two dictionaries have not fully coupled, allowing the mapping function to have much flexibility for accurate synthesis across fields.

This chapter is organized as follows. In Section 5.1, the basics of the paradigm of the sparse representation of signal is reviewed. The proposed sparse representation based ABWE approach is described in Section 5.2. The schemes explored for enhancing the proposed ABWE approach for unvoiced speech are discussed in Section 5.3. In Section 5.4, the semi-coupled dictionary based ABWE is described. The enhanced version of SCD using clustering is described in Section 5.5. In Section 5.6 the details of the experimental setup and the performance measures used are provided. The experimental results are presented in Section 5.7. Section 5.9 summarizes the paper along with suggesting direction of the future work.

5.1 Review of Sparse Representation

Sparsity is defined as having few non-zero components or having few components that are not zero. In signal processing, it is always desirable to seek compact representation of the signals. The choice of sparsity as a desired characteristic of representation of the input data can be motivated by the observation that most sensory data such as natural images may be described as the superposition of a small number of elements such as surfaces or edges. Sparse representation (SR) technique is widely used in many signal processing application such as, image de-noising [147], image compression [148], etc. Also, in recent days SR technique has gained its importance in various speech signal processing applications such as speaker verification [149], speech recognition [150] etc.

The sparse representation technique mainly consists of two stages namely: (i) the choice of dictionary and (ii) the sparse coding stage. The fundemental question in the SR technique is the choice of dictionary selection. There are two possibilities for dictionary selection i. e., either a predefined transform such as Fourier, Wavelet, etc., or from the data using some dictionary learning approaches. The major disadvantage of predefined transform is the limitation of model functions used to represent the target data. These model functions are too simple to represent the complex natural data. Where as, in dictionary learning approach, the dictionaries are

learned from the data. So the structure of data to be represented can be more accurately extracted during the learning process. The K-SVD algorithm is one among such dictionary learning approaches which is popularly used in many image processing applications. The main aim of sparse coding is to find a set of basis vectors from the chosen dictionary, such that the target vector can be represented as a linear combination of these basis vectors with minimum representation error. There exists a number of algorithms for this purpose. Among those algorithms, orthogonal matching persuit (OMP) is most commonly used algorithm and is very simple yet effective. OMP algorithm is greedy in nature using l_0 norm constraint to solve the objective function. Also, there are other approaches like least angle absolute shrinkage and selection operator (LASSO) algorithm and least angle regression (LARS) algorithm which solves the probel useing l_1 norm.

In most of the image processing applications we often need to transform images of one domain to another domain for better visualization or recognition [133, 151]. One among such transformations is to convert low sesolution image to high resolution image reffered to as image super resolution. Here we need to model the two feature spaces and the corresponding transformation to enhance the low resolution image. For this purpose, the sparse representation technique is employed and resulted in improved performances.

Sparse representation (SR) is a method to solve an over-complete system of linear equations. Sparse representation problem can be mathematically stated as: given a vector $\mathbf{x} \in \mathcal{R}^N$ (referred to as the *target*) and a matrix $\mathbf{D} \in \mathcal{R}^{N \times M}$ (referred to as the *dictionary* while its columns are referred to as the *atoms* and $M \gg N$), find a vector $\mathbf{w} \in \mathcal{R}^M$ such that

$$\min \|\mathbf{w}\|_0 \;\; \text{s. t.} \;\; \mathbf{D}\mathbf{w} = \mathbf{x} \tag{5.1}$$

If the vector $\mathbf{w}$ has the minimum ℓ_0-norm, i.e., has only a small numbers of non-zero elements, then it is termed as a sparse representation of $\mathbf{x}$. But the solution to this problem is NP hard. There are a number of techniques proposed in the literature to solve the sparse representation problem. These techniques can be classified into the family of *greedy* algorithms and *relaxation* algorithms.

The greedy algorithms attempt to construct the support one atom at a time. In other words, it provides an iterative solution to following mathematical problem,

$$\min \|\mathbf{Dw} - \mathbf{x}\|_2 \quad \text{s. t.} \quad \|\mathbf{w}\|_0 \leq \mathrm{T} \tag{5.2}$$

where T is the sparsity constraint. We have used the orthogonal matching pursuit (OMP) algorithm for sparse coding in this work.

5.1.1 Creation of dictionary for sparse representation

There are two kinds of dictionaries that are commonly used for the sparse representation purpose. The first one is referred to as the *exemplar dictionary*. It is a redundant dictionary derived by concatenating a large number of examples to produce a sparse presentation for the target. Such dictionaries are hand-created and are difficult to optimize. In contrast to that the *learned dictionary* is derived by processing the data to produce a sparse representation. In this work, we have used KSVD [132] algorithm for creating a learned redundant dictionary for sparse representation purpose. The KSVD is a generalization of the well known K-means clustering algorithm. It constructs a dictionary of K atoms that leads to the best possible representation for each of the training examples along with the specified sparsity constraint. The dictionary learning problem is represented as,

$$\min_{\mathbf{D},\mathbf{W}} \left\{ \|\mathbf{X} - \mathbf{DW}\|_2^2 \right\} \quad \text{s. t.} \quad \|\mathbf{w}_i\|_0 \leq T \quad \forall i \tag{5.3}$$

where, $\mathbf{X}$ is the set of dictionary training vectors, $\mathbf{D}$ is the learned dictionary, $\mathbf{W}$ is the set of corresponding sparse vectors, T is the sparsity constraint and training example vector index i. The dictionary learning is an iterative process and each iteration alternates between two stages: sparse coding and dictionary update.

5.2 ABWE using sparse representation

It is well known that the Fourier based dictionaries (DFT, DCT, etc.) have orthogonal bases and those have highly narrowband spectral characteristics. As a result of this, although

these dictionaries happen to produce very accurate representation of speech signals but their representations are neither sparse nor have much similarity for WB and NB speech. In contrast, if we create a dictionary that produces a sparse representation for speech signals, then it is more likely to produce similar representations for WB and NB speech. Motivated by this hypothesis, we have explored a very simple approach to ABWE in which the sparse representation of the NB speech obtained with respect to a NB dictionary are applied to a corresponding WB dictionary to achieve the reconstruction of HB information for the given NB signal.

5.2.1 Proposed SR-ABWE approach

For learning the dictionary for sparse modelling purpose, a development data is derived from the training set of WSJCAM0 corpus [152] which includes both male and female speakers. The selected speech signals are segmented into frames of size 20 ms with an overlap of 5 ms between frames. A total of 0.12 million frames are collected and the resulting data matrix is then used in dictionary learning using KSVD algorithm for the sparse representation of speech frames. In dictionary learning, 50 iterations of KSVD are used.

For sparse modelling, first we have created a single dictionary having 4000 atoms on WB speech. The atoms of this WB dictionary are first decimated by a factor of 2 and then interpolation by a factor of 2 to derive a dictionary for the sparse representation of the corresponding upsampled narrowband (NBI) signal frames. The atoms of so derived NBI dictionary have a *one-to-one* mapping to those of WB dictionary. The sparse representation of NBI target frames obtained with NBI dictionary are then associated with WB dictionary to synthesize a signal having enhanced HB information. For single dictionary case, the plot of the bandwidth enhanced magnitude spectra for NBI frame obtained using the proposed ABWE approach for a voiced (/aa/) and an unvoiced (/s/) frames are shown in Figure 5.1(a) and Figure 5.1(c), respectively. For contrast purpose, the magnitude spectra for the sparse representation of WB frame with WB dictionary and the original WB signal are also shown in figure. The corresponding sparse codes for NBI/WB data are shown in Figure 5.1(b) and Figure 5.1(d). On comparing with the original signal spectra, the proposed ABWE modelling approach appears

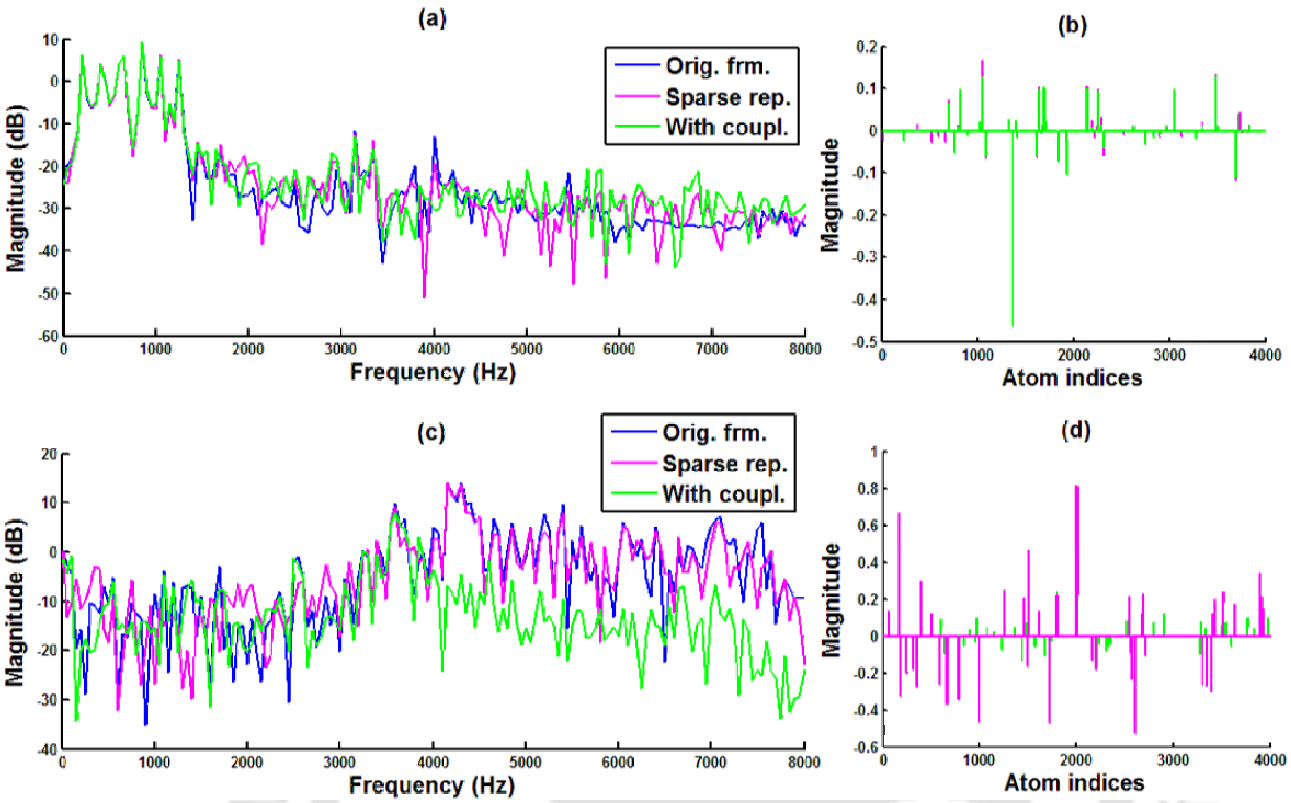

Figure 5.1: Panels (a) and (c) show the reconstructed spectra using the proposed ABWE approach for a voiced (/aa/) and a unvoiced (/s/) frames of speech, respectively. For contrast purpose, the spectra for the sparse representation of WB frame and the original frame are also shown. Panels (b) and (d) show the corresponding sparse codes obtained with WB and NBI dictionaries. A single dictionary is used for sparse coding of both voiced and unvoiced segments.

working somewhat for voiced case, but for unvoiced case it does not appear to be effective enough.

On noting the fact that the characteristics of voiced and unvoiced speech differ significantly, so a single dictionary may not be very effective. To improve the modelling, it was decided to learn separate dictionaries for voiced and unvoiced speech cases. For judging the improvement in the proposed ABWE modelling with separate voiced and unvoiced KSVD learned dictionaries, the magnitude spectra for the same voiced and unvoiced frames are shown in Figure 5.2. On comparing Figure 5.1 and Figure 5.2, a considerable improvement in SR-ABWE modelling can be noted with the use of separate dictionaries, in particular for the unvoiced case. Though for unvoiced case, the indices in NBI and WB sparse codes still differ but the enhanced spectra

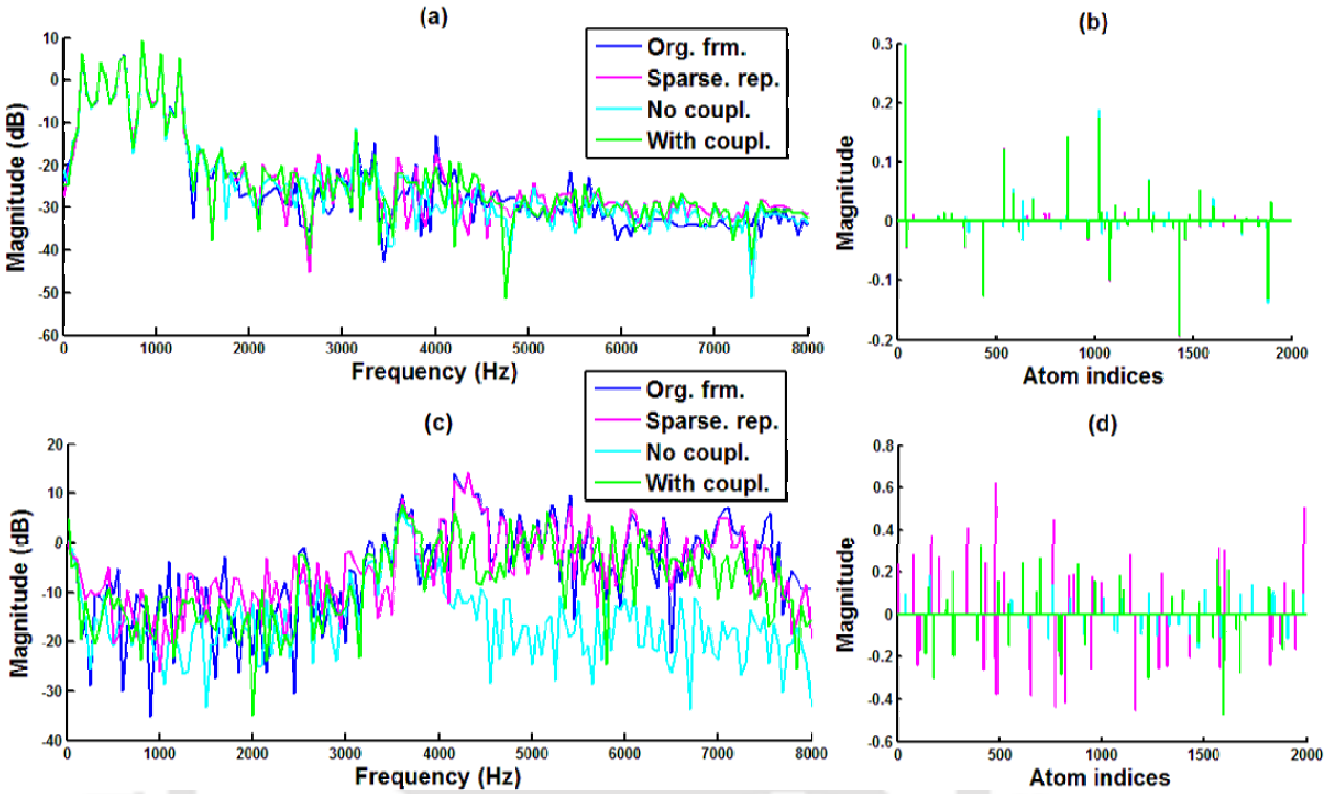

Figure 5.2: Improvement in the modelling of the proposed ABWE approach with the use of separate voiced and unvoiced dictionaries in sparse coding. The example frames and the layout of panels are identical to that of Figure 5.1. Separate dictionaries are used for voiced and unvoiced cases.

show better match with the original WB spectra.

Like other existing ABWE approaches, in the proposed approach too the given NB speech is retained without any modification while the estimated HB speech is added to the given NB speech with appropriate amplitude scaling. The detailed block diagram of the proposed sparse representation based ABWE approach in shown in Figure 5.3. Further to highlight that the creation of coupled NBI and WB dictionaries is critical for the proposed ABWE approach. Simply the sparse coding of NBI speech data with WB dictionary would lead to poor recreation of HB information especially for unvoiced case. To verify this fact, the enhanced spectra for NBI frame obtained by sparse coding with WB dictionary are also given in Figure 5.1(a) and Figure 5.1(c) for voiced and unvoiced cases, respectively. Note that very poor enhancement is obtained without any coupling for unvoiced case.

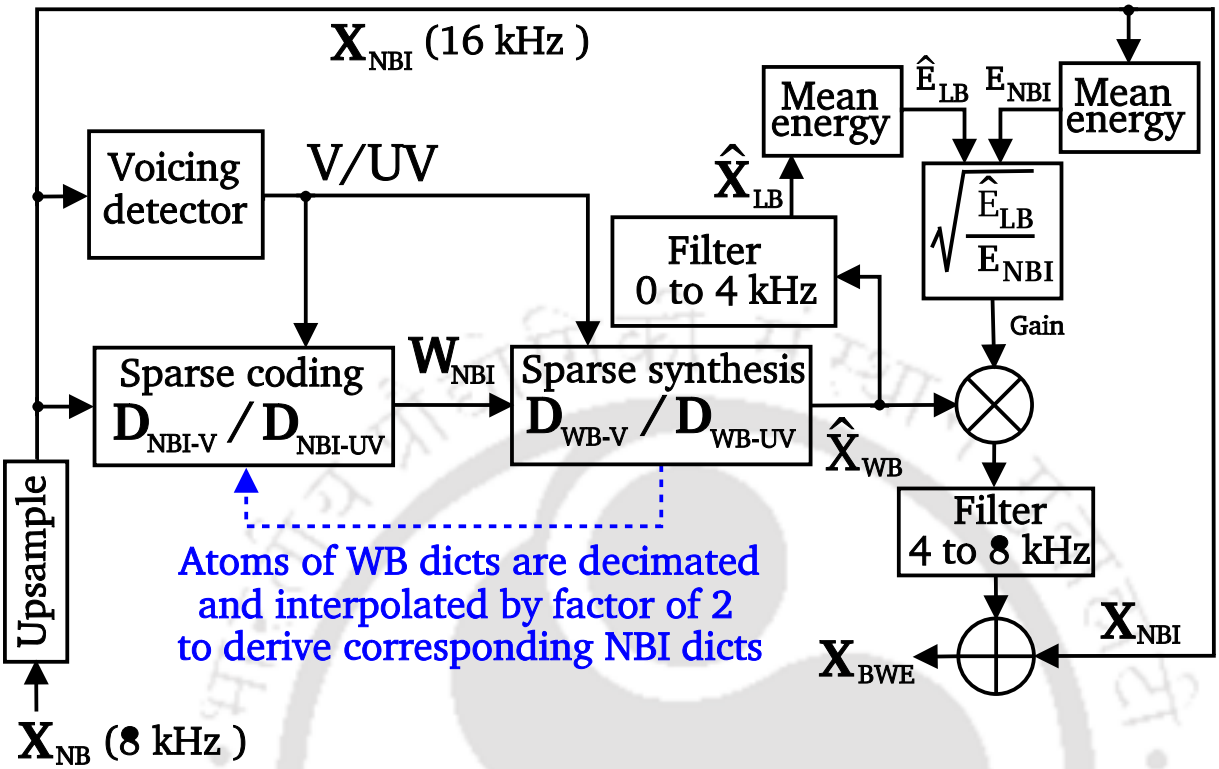

Figure 5.3: The complete block diagram of the proposed sparse representation based ABWE approach.

5.3 Enhancements in proposed SR-ABWE approach

The proposed approach seems quite effective for the voiced speech case, but for the unvoiced speech case it definitely requires improvement. In the following, the enhancements explored in the proposed SR-ABWE approach to improve the quality of the bandwidth extension for the unvoiced speech while making no changes for the voiced speech.

5.3.1 Linear transformation of NBI sparse coefficients

It is well known that with bandwidth reduction of speech signal, a significant loss of HB spectral information occurs in particular for unvoiced case. On account of this, on sparse coding of NBI frame with NBI dictionary, the effective representation could be achieved by those atoms of dictionary which do not correspond to the ones having significant HB information in WB dictionary. Consequently, on using the resulting sparse code for synthesis with WB dictionary,

the HB spectral information is not regenerated in effective manner. This fact can be verified by studying the vastly different sparse coding as well as deficiency in the spectral information produced for WB and NBI speech frame corresponding to unvoiced case as shown in Figure 5.2.

As a first recourse we explored a linear transformation to address the significant differences of the sparse coding for NBI and WB cases. Let $\mathbf{W}_{\text{NBI}}$ and $\mathbf{W}_{\text{WB}}$ denote the sparse code matrices for NBI and WB cases for the unvoiced frames in the training data, then a least squares (LS) based linear transformation $\mathbf{T}_{\text{LS}}$ is estimated as

$$\mathbf{T}_{\text{LS}} = \mathbf{W}_{\text{WB}}\mathbf{W}_{\text{NBI}}^{T}(\mathbf{W}_{\text{NBI}}\mathbf{W}_{\text{NBI}}^{T})^{-1} \tag{5.4}$$

For unvoiced case, an adapted sparse codes matrix given the target NBI sparse code matrix is derived as

$$\widetilde{\mathbf{W}} = \mathbf{T}_{\text{LS}}\mathbf{W}_{\text{NBI}} \tag{5.5}$$

Obviously, on linear transformation, the sparsity in resulting vector is lost and a large number of atoms acquire non-zero coefficient value. Therefore, in final synthesis with WB dictionary for bandwidth extension, a synthetic sparsity is enforced by retaining the coefficients of the highest magnitude atoms totaling the chosen sparsity value (T) while setting the rest of them to zero.

The attempted adaptation of the NBI sparse codes for unvoiced frames is expected to produce more enhanced HB information than the default case. To assess its efficacy, the spectral profile of the atoms involved in WB sparse code, NBI sparse code and after the linear transformation of NBI sparse code for an unvoiced frame are shown in Figure 5.4(a), 5.4(b) and 5.4(c), respectively. We can note that after linear transformation of NBI sparse codes, the selected atoms are found to possess somewhat more HB information.

5.3.2 Lookup constrained linear transformation

Though the simple adaptation of NBI sparse codes appears to help in modelling for unvoiced case, but there is scope for further improvement. We hypothesize that if the information about the atoms which potentially have significant HB information can be utilized while selecting the

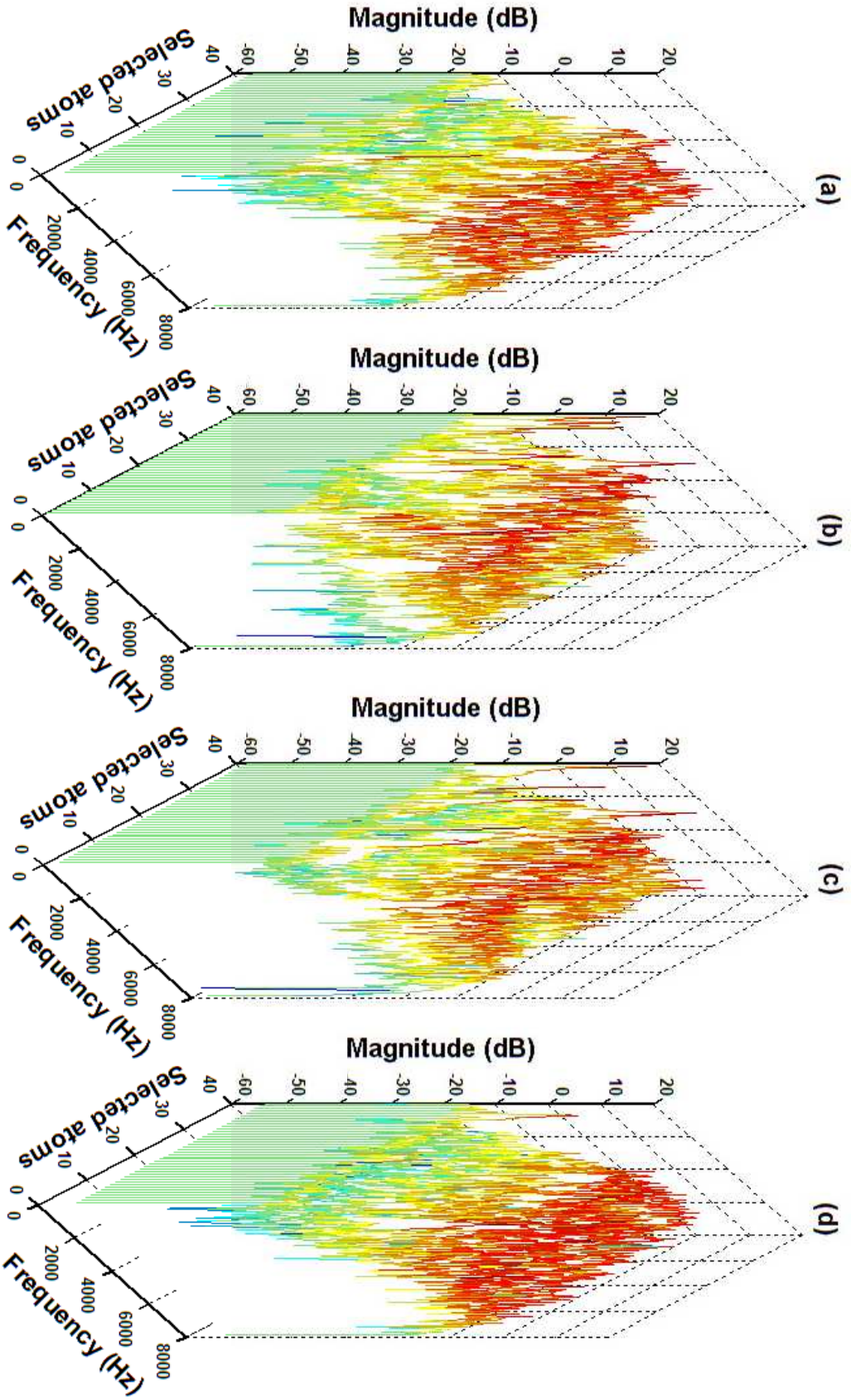

Figure 5.4: Plots showing the spectral profile of the atoms involved in the sparse representation of an unvoiced frame for (a) WB sparse code (b) NBI sparse code (c) top 40 of the linear transformed NBI sparse code and (d) lookup table based WB sparse code.

significant atoms after linear transformation, the quality of bandwidth extension for unvoiced case could be further boosted. To provide this additional information, a *lookup-table* is created by preserving the representative WB sparse codes of the unvoiced frames in the development data set. This lookup table is indexed by the code vectors of a VQ developed on MFCC features corresponding the NBI unvoiced speech frames in the development data set. The procedure for the lookup-table creation is elaborated in Algorithm 1.

Given an unvoiced NBI speech frame, the corresponding MFCC feature vector is also computed which is used to find the nearest (in Euclidean sense) code vector in the VQ codebook. Based on the index of the code-vector, a putative WB sparse code is noted from the lookup table. Though so derived putative WB sparse code may not be optimal for the given NBI frame but it does provide the knowledge about the atoms potentially having the HB information. To use this knowledge, a binary mask is constructed selecting those atoms that are indicated in the putative WB sparse code. This binary mask is then multiplied with the linear transformed NBI sparse code for the given frame to derive the enhanced sparse code for synthesis with WB dictionary for ABWE purpose. The block diagram of the procedure employed for obtaining the enhanced WB sparse code for unvoiced cases is shown in Figure 5.5.

Figure 5.4(d) shows the spectral profile of the atoms obtained by the lookup approach. On comparing that with other cases, we can note that a significant increase in energy for the HB region is achieved.

5.4 Semi-Coupled Dictionary based ABWE

The detailed block diagram of the semi-coupled dictionaries based ABWE (SC-ABWE) approach is shown in Figure 5.6. As shown in the block diagram, speech frames are first classified using a V/UV detector. The KSVD dictionary of sparsity one as discussed in Sec. 5.5 is used for unsupervised classification of speech frames. $\mathbf{C}_{V-k}$ and $\mathbf{C}_{UV-k}$ represents the voiced and unvoiced class for index k, respectively. The creation of semi-coupled dictionaries is elaborated in Sec. 5.4.1. Similar to other existing sparse representation based ABWE approaches, the given NB speech is retained without any modification and the estimated HB speech is added to

Algorithm 1 Procedure for creation of the lookup-table used for adapting the sparse codes of the unvoiced speech frames

Given: Wideband (WB) development data corresponding to the unvoiced region in the speech only
Given: The WB data is segmented into 20 ms frames with 5 ms overlap and stacked column wise into a matrix
Step 1: Learn a dictionary $\mathbf{D}_{\text{WB-UV}}$ having 2000 columns with sparsity constraint of 10 using KSVD algorithm on WB data
Step 2: Sparse code the columns of WB data matrix on $\mathbf{D}_{\text{WB-UV}}$ with sparsity constraint of 40 using OMP algorithm and store the sparse codes into a matrix
Step 3: Each column of WB data matrix is both decimated and interpolated by a factor of 2 to get the narrowband interpolated (NBI) data matrix
Step 4: For each of the column of NBI data matrix, compute the 39-dimensional MFCC features which includes 13 static features with their delta and delta-delta appended
Step 5: Vector quantize the NBI MFCC data matrix into a 1024 size codebook
Step 6: For each of code vectors in the codebook, find the index of the NBI data vector that is nearest (in Euclidean sense) to that code vector
Step 7: Create a lookup-table that maps the MFCC code vectors for NBI data to the WB sparse code corresponding to the index of the nearest data vector found in Step 6

the given NB speech with appropriate amplitude scaling. Given the NBI sparse codes of speech data, using the pre-learned mapping function, the WB sparse codes are estimated. Using the estimated wideband sparse codes and the wideband dictionary, the WB speech is synthesized.

5.4.1 Semi-coupled dictionary algorithum

Semi-coupled dictionary algorithm is initially proposed in [133] for photo-sketch synthesis. In this work we are discussing the creation of SC dictionaries in the context of ABWE. Let X and Y denote the training datasets formed by the speech frame pairs of NBI and WB speech signals. The objective is to minimize the energy function given below to find the desired semi-coupled dictionaries and the desired mapping function.

$$\begin{aligned}&\min_{\{\mathbf{D}_x,\mathbf{D}_y,f(.)\}}\{E_{data}(\mathbf{D}_x,\mathbf{X})+E_{data}(\mathbf{D}_y,\mathbf{Y})\\&+\gamma E_{map}(f(\mathbf{\Lambda}_x,\mathbf{\Lambda}_y))+\lambda E_{reg}(\mathbf{\Lambda}_x,\mathbf{\Lambda}_y,f(.),\mathbf{D}_x,\mathbf{D}_y)\}\end{aligned} \tag{5.6}$$

where $E_{data}(.,.)$ is the data fidelity term to represent data description error, $E_{map}(.,.)$ is the mapping fidelity term to represent the mapping error between the coding coefficients of two spaces, and E_{reg} is the regularization term to regularize the coding coefficients and mapping. Note that in this model, the coding coefficients of $\mathbf{X}$ and $\mathbf{Y}$ over $\mathbf{D}_x$ and $\mathbf{D}_y$ will be related

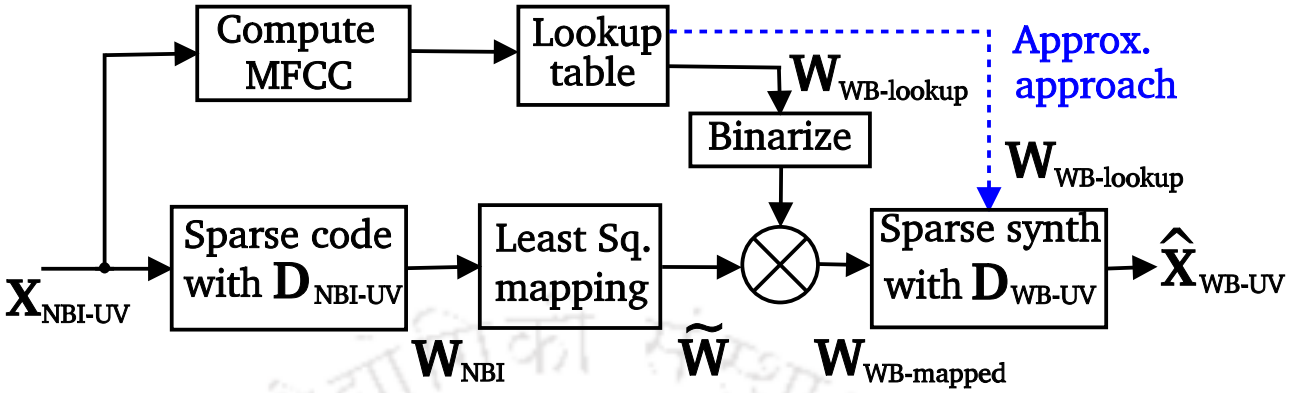

Figure 5.5: Block diagram of lookup-constrained linear transformation approach explored for addressing the mismatch in sparse codes obtained with using NBI and WB dictionaries for unvoiced case. The alternative of using the looked-up sparse code directly for ABWE is denoted as the "approximate approach".

by a mapping $f(.)$. The two dictionaries ($\mathbf{D}_x$ and $\mathbf{D}_y$) and the mapping function $f(.)$ will be jointly optimized.

If the mapping $f(.)$ is assumed to be linear, then the framework in Eq. 5.6 can be turned into the following dictionary learning and ridge regression problem:

$$\begin{aligned} &\min_{\{\mathbf{D}_x,\mathbf{D}_y,\mathbf{W}\}} \|\mathbf{X}-\mathbf{D}_x\mathbf{\Lambda}_x\|_F^2+\|\mathbf{Y}-\mathbf{D}_y\mathbf{\Lambda}_y\|_F^2 \\ &+\gamma\|\mathbf{\Lambda}_y-\mathbf{W}\mathbf{\Lambda}_x\|_F^2+\lambda_x\|\mathbf{\Lambda}_x\|_1+\lambda_y\|\mathbf{\Lambda}_y\|_1+\lambda_w\|\mathbf{W}\|_F^2 \\ &\text{s.t. } \|\mathbf{d}_{x,i}\|_{l_2}\leq 1, \|\mathbf{d}_{y,i}\|_{l_2}\leq 1, \forall i \end{aligned} \tag{5.7}$$

where $\gamma, \lambda_x, \lambda_y, \lambda_W$ are regularization parameters to balance the terms in the objective function and $\mathbf{d}_{x,i}$, $\mathbf{d}_{y,i}$ are the atoms of $\mathbf{D}_x$ and $\mathbf{D}_y$, respectively. The objective function in Eq. 5.7 is not jointly convex to $\mathbf{D}_x$, $\mathbf{D}_y$, $\mathbf{W}$. However, it is convex with respect to each of them if others are fixed. Therefore, we can design an iterative algorithm to alternatively optimize the variables.

5.4.2 Training

To tackle the energy minimization in Eq. 5.7, the objective function is separated into 3 subproblems, namely sparse coding for training samples, dictionary updating and mapping updating. First, the mapping $\mathbf{W}$ and dictionary pair are to be initialized. $\mathbf{W}$ can be simply

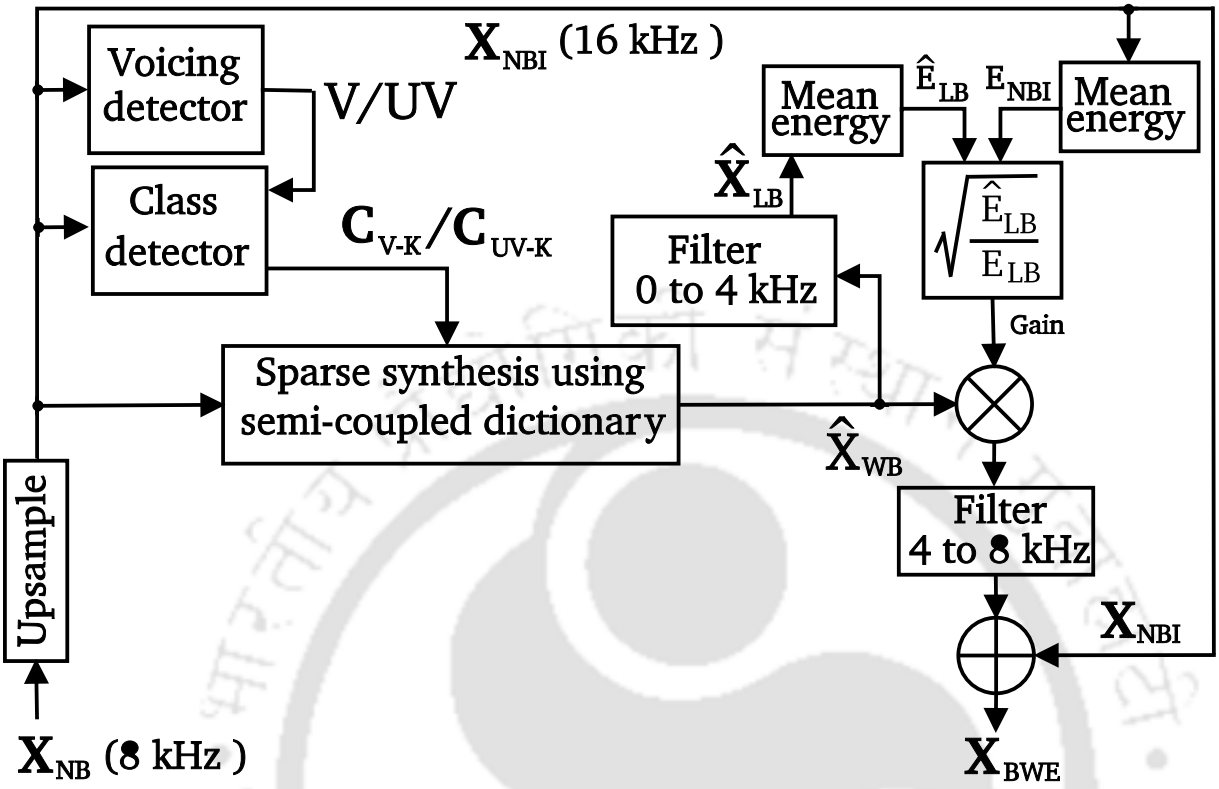

Figure 5.6: The complete block diagram of the proposed sparse representation based ABWE approach.

initialized as the identity matrix. Using l_1-minimization, the sparse codes $\mathbf{\Lambda}_x$ and $\mathbf{\Lambda}_y$ can then be obtained. Note that mapping by $\mathbf{W}$ is assumed to be linear, and the bidirectional transform learning strategy can be adopted to learn transforms from $\mathbf{\Lambda}_x$ to $\mathbf{\Lambda}_y$ and from $\mathbf{\Lambda}_y$ to $\mathbf{\Lambda}_x$ simultaneously.

With some initialization of $\mathbf{W}$ and dictionary pair $\mathbf{D}_x$ and $\mathbf{D}_y$, it can calculate the sparse coding coefficients $\mathbf{\Lambda}_x$ and $\mathbf{\Lambda}_y$ as follows:

$$\begin{aligned} &\min_{\{\mathbf{\Lambda}_x\}} \|\mathbf{X} - \mathbf{D}_x\mathbf{\Lambda}_x\|_F^2 + \gamma\|\mathbf{\Lambda}_y - \mathbf{W}_x\mathbf{\Lambda}_x\|_F^2 + \lambda_x\|\mathbf{\Lambda}_x\|_1 \\ &\min_{\{\mathbf{\Lambda}_y\}} \|\mathbf{Y} - \mathbf{D}_y\mathbf{\Lambda}_y\|_F^2 + \gamma\|\mathbf{\Lambda}_x - \mathbf{W}_y\mathbf{\Lambda}_y\|_F^2 + \lambda_y\|\mathbf{\Lambda}_y\|_1 \end{aligned} \tag{5.8}$$

Eq. 5.8 is a multi-task lasso problem. Many l_1-optimization algorithms can solve it effectively. In this work, least-angle regression (LARS) method is chosen as the l_1-optimization

method for its efficiency and stability. With $\mathbf{\Lambda}_x$ and $\mathbf{\Lambda}_y$ fixed, dictionary pair $\mathbf{D}_x$ and $\mathbf{D}_y$ can be updated as follows:

$$\begin{aligned} &\min_{\{\mathbf{D}_x,\mathbf{D}_y\}} \|\mathbf{X}-\mathbf{D}_x\mathbf{\Lambda}_x\|_F^2+\|\mathbf{Y}-\mathbf{D}_y\mathbf{\Lambda}_y\|_F^2 \\ &\text{s.t. } \|\mathbf{d}_{x,i}\|_{l_2}\le 1, \|\mathbf{d}_{y,i}\|_{l_2}\le 1 \end{aligned} \tag{5.9}$$

Eq. 5.9 is a quadratically constrained quadratic program problem (QCQP) and a one-by-one update strategy is adopted to solve it.

$$\min_{\{\mathbf{W}\}} \|\mathbf{\Lambda}_y-\mathbf{W}_x\mathbf{\Lambda}_x\|_F^2+(\lambda_W/\gamma).\|\mathbf{W}\|_F^2 \tag{5.10}$$

With dictionary and coding coefficients fixed, it can then update the mapping $\mathbf{W}$:

$$\mathbf{W}=\mathbf{\Lambda}_y\mathbf{\Lambda}_x^T(\mathbf{\Lambda}_x\mathbf{\Lambda}_x^T+(\lambda_W/\gamma).\mathbf{I})^{-1} \tag{5.11}$$

where $\mathbf{I}$ is an identity matrix.

With SCDL, it can learn the dictionary pair $\mathbf{D}_x$ and $\mathbf{D}_y$ on which the sparse coding coefficients of two spaces have stable bidirectional linear transformations. In Section 5.5 we can further enhance its stability by clustering samples into several clusters.

The SCDL learning algorithm is summarized as in Algorithm 2.

Algorithm 2 Semi-Coupled Dictionary Learning

Input: Training datasets $\mathbf{X}$ and $\mathbf{Y}$ of two speech cases, namely, NBI and WB. Each corresponding pair indicates the same speech. Initial dictionary pair $\mathbf{D}_x$ and $\mathbf{D}_y$, and initial mapping $\mathbf{W}_x$ and $\mathbf{W}_y$.
For each iteration **Until** convergence:
For each cluster

(i) Fix other variables, update $\mathbf{\Lambda}_x$ and $\mathbf{\Lambda}_y$ by sparse coding in Eq. 5.8.

(ii) Fix other variables, update $\mathbf{D}_y$ and $\mathbf{D}_y$ in Eq. 5.9.

(iii) Fix other variables, update $\mathbf{W}_x$ and $\mathbf{W}_y$ in Eq. 5.10.

Update clustering index of each pair as described in Subsec. 5.5
Output: $\mathbf{D}_y$, $\mathbf{D}_y$, $\mathbf{W}_x$ and $\mathbf{W}_y$

5.4.3 Synthesis

After learning the dictionaries $\mathbf{D}_x$ and $\mathbf{D}_y$ and the linear mapping $\mathbf{W}$, for a given speech $\mathbf{x}$ in NBI type, we can easily convert it into a WB speech by solving the following optimization:

$$\begin{aligned}\min_{\{\alpha_{x,i},\alpha_{y,i}\}} &\|\mathbf{x}_i - \mathbf{D}_x\alpha_{x,i}\|_F^2 + \|\mathbf{y}_i - \mathbf{D}_y\alpha_{y,i}\|_F^2 \\ &+\gamma\|\alpha_{y,i} - \mathbf{W}_y\alpha_{x,i}\|_F^2 + \gamma\|\alpha_{x,i} - \mathbf{W}_x\alpha_{x,i}\|_F^2 \\ &+\lambda_x\|\alpha_{x,i}\|_1 + \lambda_x\|\alpha_{y,i}\|_1 \end{aligned} \tag{5.12}$$

where $\mathbf{x}_i$ is a frame of NBI and $\mathbf{y}_i$ is the corresponding frame in the intermediate estimate of y to be synthesized. Eq. 5.12 can be solved by alternatively updating $\alpha_{x,i}$ and $\alpha_{y,i}$. Finally, each frame of $\mathbf{y}$ can be reconstructed as:

$$\hat{\mathbf{y}}_i = \mathbf{D}_y\hat{\alpha}_{y,i} \tag{5.13}$$

After all the frames are estimated, the estimation of the desired speech y can then be obtained.

In this synthesis method, an initial estimation of y is needed. Depending on the problem, different strategies can be adopted to initialize y. In the problem of ABWE synthesis, we can first code $\mathbf{x}_i$ on $\mathbf{D}_x$ for coding vector $\alpha_{x,i}$ and then initialize $\mathbf{y}_i$ as $\mathbf{D}_y\mathbf{W}\alpha_{x,i}$.

The SCDL synthesis process is summarized as described in Algorithm 3.

Algorithm 3 SC-ABWE Synthesis

Input: Test speech frame $\mathbf{x}_i$, well trained dictionary pair $\mathbf{D}_x$ and $\mathbf{D}_y$, the learnt mapping $\mathbf{W}_x$ and $\mathbf{W}_y$ for two cases, namely NBI and WB.
Initialization: Initialize $\mathbf{y}_i$ as discussions in subsec 5.4.3. Initialize clustering index of the frame.
For each iteration **Until** convergence:

(i) optimization as Eq. 5.12

(ii) Update $\mathbf{y}$ as the synthesis in Eq. 5.12.

Output: Synthesized ABWE speech frame $\mathbf{y}$.

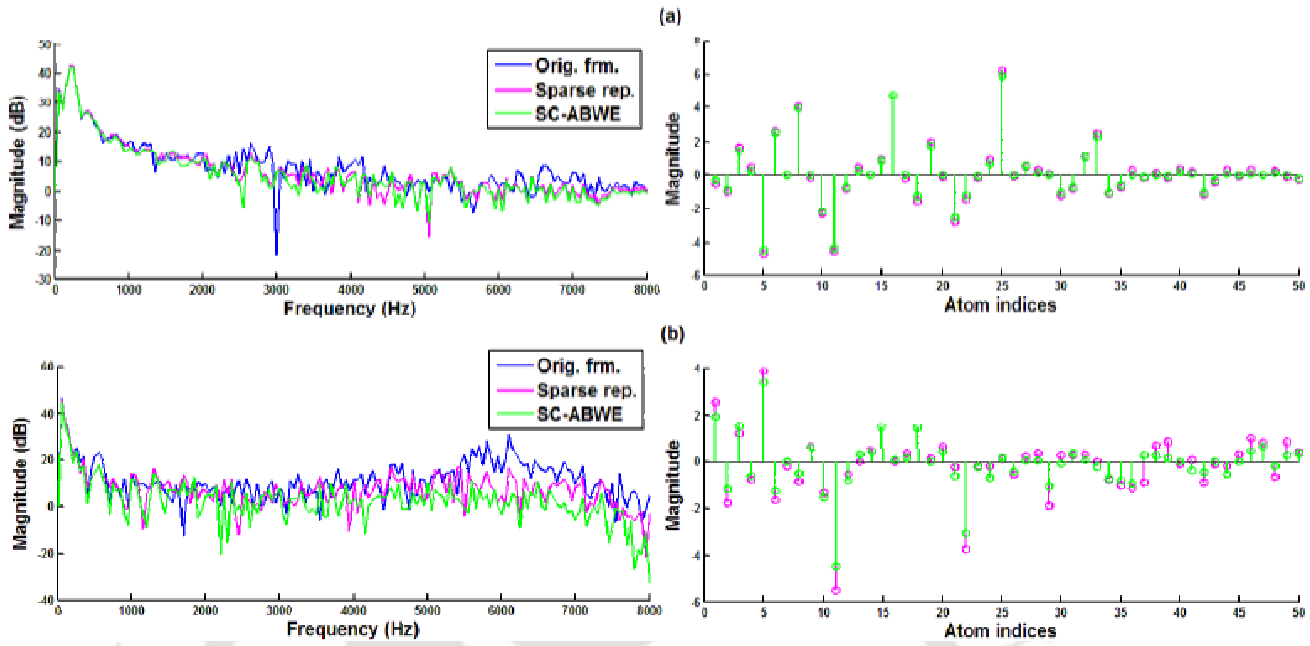

Figure 5.7: Panels (a) and (b) show the reconstructed spectra using the proposed SC-ABWE approach for a voiced and a unvoiced frames of speech, respectively. For contrast purpose, the spectra for the sparse representation of WB frame and the original frame are also shown. Adjacent panels show corresponding sparse codes obtained with WB and transformed NBI codes.

5.5 Clustering based SCDL ABWE

Due to the diverse nature of speech of different phone classes, learning only one pair of dictionaries and an associated linear mapping function is often not enough to cover all variations of speech. We therefore chose to cluster the data by NBI KSVD-dictionary of sparsity one. NBI KSVD-dictionary is produced from NBI and WB joint dictionary learning process. NBI and WB joint dictionary is learnt by concatenating both NBI and WB frames. We trained 64 atoms KSVD dictionary for both voiced and unvoiced class. Voiced and unvoiced classification carried out using *FXRAPT* function of Voice box toolbox.

The class of NBI test speech frame is found using the sparse coded NBI KSVD-dictionary with sparsity one. The index of the atom with highest correlation to the test frame is taken as the clustering label.

5.6 Experimental Setup and Performance Measures

For the evaluation of the proposed approach, the speech data is taken from WSJCAM0 [152] database commonly used for developing the automatic speech recognition systems. In WSJCAM0 database, the partitioning of training and testing sets are available. For learning the dictionary for the sparse representation, a development set was created by randomly choosing one file each from 92 speakers available in the training set. The evaluation of the developed methods is done on the "5k-test set" consisting of 368 files from 20 speakers. All speech data is segmented into 20 ms rectangular windowed frames keeping 5 ms overlap between the frames. To create separate dictionaries for voiced and unvoiced speech, the overall development data consisting of 1.2×10^5 frames is marked using *FXRAPT* function of Voicebox [153]. That resulted in 76×10^3 and 44×10^3 frames for voiced and unvoiced speech cases, respectively. While learning the dictionary with KSVD algorithm, a sparsity value of 10 is used whereas a sparsity value of 40 is employed for sparse coding the speech data for ABWE. These sparsity values are chosen based on experimentation. For the purpose of creation of the lookup table, the unvoiced NBI speech frames of the development data are also parameterized into 39-dimensional MFCC features which include 13 base features and their corresponding velocity and acceleration coefficients.

The Speech Quality measures are described in Appendix A.

5.7 Experimental Results and Discussion

For contrast purpose, the performance for a conventional ABWE method [146] is also evaluated. It is a traditional method and employs the source-filter model of speech. In the contrast method, the information in NB and HB portions of speech is captured in LSF domain. A joint NB-HB (10-6 dimensional) LSF features are computed and modelled using a 32 component Gaussian mixture model having a full covariance matrix. Given the NB information, the missing HB information is estimated following minimum mean square error (MMSE) criterion. Therefore the estimated HB speech is then added with suitable scaling to the given NB speech to produce the bandwidth extended speech.

Table 5.1: *Performances for the proposed SR-ABWE approach in the default case and including those with the global linear transformation (LT), the approximate approach (use of looked-up WB sparse code) and the lookup-constrained global LT applied for unvoiced (UV) frames only.*

Method	Enhancement applied for UV frames	Measures			
		d_{LSD}	segSNR	d_{LLR}	PESQ
SR-ABWE	Default	11.04	9.62	0.64	4.46
	Linear transform	9.85	10.44	0.68	4.46
	Approx. approach	9.81	10.49	**0.56**	4.46
	Lookup-constrained LT	**9.64**	**10.53**	0.64	**4.46**
Conventional-ABWE (LSF based)		11.18	8.53	0.60	4.14

The performances for the contrast method and the proposed SR-ABWE approach along with that of its enhancements are given in Table 5.1. On comparing with the contrast approach, we note that the proposed SR-ABWE approach in the default case has resulted in better performances for all measures expect for slight degradation in case of d_{LLR}. For addressing the weakness of the proposed ABWE framework in the unvoiced case, both the linear adaptation of NBI sparse codes and lookup-constrained linear adaptation have resulted significant improvement in performance for almost all measures. Interestingly, the use of lookup based putative WB sparse code without linear transformation is also found to be quite effective, though not for all measures, for addressing the poor modelling issue for the unvoiced case.

To achieve optimal performance in case of semi-couple dictionary by tuning, the different results obtained for the case of coupled dictionaries are given in Table 5.2 . The couple dictionaries case involve joint training of the dictionaries, where as the semi-coupled also involves learning the mapping function. The corresponding results for the semi-couple dictionaries are also tabulated in Table 5.3. As it can be observed, the semi-coupled dictionary case provides the best performance of 8.64 for d_{LSD} when 20 atoms are used.

Table 5.2: *Performances for the Coupled-ABWE, SC-ABWE Performances with different number of atoms.*

Method	No of atoms	Measures			
		d_{LSD}	segSNR	d_{LLR}	PESQ
Copuled-ABWE	20	**8.78**	**11.14**	0.92	4.45
	50	8.93	10.67	0.79	4.45
	100	10.26	9.71	**0.62**	4.45
Method	No of atoms	Measures			
		LSD	segSNR	LLR	PESQ
SC-ABWE	20	**8.64**	**11.25**	0.95	4.41
	50	8.73	10.76	0.80	4.43
	100	9.13	10.43	**0.71**	4.43

Table 5.3: *Performances for the proposed SC-ABWE approach and the lookup-constrained global LT applied for unvoiced (UV) frames only.*

Method	Enhancement applied for UV frames	Measures			
		d_{LSD}	segSNR	d_{LLR}	PESQ
SR-ABWE	Lookup-constrained LT	9.64	10.53	0.64	**4.46**
SC-ABWE		**8.64**	**11.25**	0.95	4.41
Conventional-ABWE (LSF based)		11.18	8.53	**0.60**	4.14

5.8 Application of Sparse Representation based ABWE in Children's Speech ASR

It is well known that the acoustic attributes between adults' and children's speech differ significantly. As a result of that when children's speech recognition is performed on adults' speech trained ASR systems a high degradation in the recognition performance is noted. Further, when NB speech data used for developing the ASR system instead of WB data, the recognition performances for both adults' and children's undergo significant degradation but its extent is larger for the recognition of children's mismatched speech condition, i.e., with respect to adults' speech trained acoustic models. This behavior is attributed to greater loss of spectral information in higher band in case of children's speech unlike that in adults' speech with reduction in the bandwidth of acquired speech signals. The effectiveness of the SR-ABWE approach and modifications explored in this work seem to be encouraging in the speech domain. It would be interesting to evaluate whether the bandwidth enhancement explored are also helpful in bridging the gap between recognition performance for WB and NB speech in case of children's mismatched speech recognition. Towards this purpose, a conventional context-dependent hidden Markov model (CD-HMM) based ASR system is developed using HTK toolkit [72]. The parameters of the ASR system is learned using the adults' (male and female) speech data from WSJCAM0 [152] speech corpus.

The ASR systems in this work are developed by following the procedure described in the earlier work [154]. The speech analysis is carried out using a Hamming window of length 25 ms, frame rate of 100 Hz and a pre-emphasis factor of 0.97. The MFCC base feature vector of 13-dimension is computed from a 21-channel mel filterbank using the HTK toolkit. The first- and second-order temporal derivatives, computed over a span of 5 frames, are appended to the MFCC base features, resulting a final 39-dimensional feature vector, henceforth referred to as the MFCC features. Cepstral mean subtraction is also applied to all features during training and testing.

5.8.1 Speech database

The evaluation of the proposed ABWE approaches is done on the speech data obtained from PFSTAR [9] speech corpus which is commonly used for evaluating the ASR performance of children's speech. In the PFSTAR corpus, the speech data is partitioned into training and test sets. For learning the dictionaries, a smaller development set is created by randomly selecting 127 utterances such that at least one speech file is considered for each of 80 speakers (male and female) from the existing training set. Similarly, for evaluating the ABWE approaches a smaller size test set is created by randomly selecting 115 utterances such that at least one speech file is considered for each of 60 speakers (male and female) from the existing test set. The selected speech data is analyzed into frames of 20 ms length keeping a frame-shift of 5 ms. For creating the separate dictionaries for modelling voiced, and unvoiced cases, the speech frames in the development and the test data are labeled into voiced/unvoiced using appropriate functions available in the *voicebox* [153], a commonly used MATLAB based speech toolbox. That resulted in a total of $382,148$ voiced and $329,882$ unvoiced frames in the training set, whereas the test data contained a total of $339,995$ voiced and $317,449$ unvoiced frames.

5.8.2 System parameter tuning

The dictionaries in the default and the proposed single clustered approaches are learned using KSVD algorithm with a sparsity value of 10, number of dictionary atoms as 1000, iterations 50. During the sparse coding stage, the representation sparsity value of 50 is considered.

For the proposed sub-class dictionaries learning approach, we experimentally tuned the number of sub-class dictionary atoms to be 20 and the number of clusters in each of the broad speech class to be 64. The number of iterations for each of the sub-class dictionary learning is kept same as that of the previous single clustered approach. In this approach, for the purpose of sub-class dictionaries learning and sparse coding techniques, all the atoms of the sub-class dictionary are considered.

5.8.3 Results and discussion

The effectiveness of the proposed modifications in the existing SR-ABWE approach are evaluated in terms of different speech quality measures and the results of the same are given in Table 5.4. Further, the sub-classification within each broad speech class has also helped in improving in the quality of the bandwidth enhancement. The significant improvements are noted for d_{LSD} and d_{LLR}, although a slight degradation is noted for segSNR. Note that PESQ measure turned out to be more or less same for both cases. This appeared some what strange which motivated us to try finding the effect of missing higher band information on PESQ. For the NBI speech PESQ score turned out to be 4.49 whereas the same for WB speech turned out as 4.5, thus higher band spectral information does not appear to affect the score. As in the ABWE processing, the given NB spectral information does get affected slightly while adding the higher band information which accounts for small degradation noted in PESQ scores for ABWE methods.

Table 5.4: *Performances for the No enhancement, SR-ABWE V/UV 1-cluster and V/UV 64-cluster. The quality measures are also computed for simply upsampled narrowband speech and the same is referred to as 'No enhancement'.*

System	d_{LSD}	d_{LLR}	segSNR	PESQ
No enhancement	13.99	4.03	14.62	4.49
V/UV, 1-cluster	9.08	1.69	**10.00**	4.42
V/UV/, 64-clusters	**8.29**	**1.26**	12.83	4.43

The ASR performances for the default and the proposed approaches are measured and noted in Table 5.5. It has been observed that the proposed clustering based dictionary learning technique for SR-ABWE approach using single cluster results in improved ASR performance when compared to that of non-enhanced speech data. However, the 64 cluster case does not improve the ASR performance, even though the objective measures showed improvement. This needs further exploration.

Table 5.5: *Recognition performances for children's speech (CH) test sets under mismatched condition having narrowband (NB), wideband (WB), SR-ABWE V/UV 1-cluster and V/UV 64-cluster.*

System	NB	WB	V/UV 1-cluster	V/UV 64-cluster
WER(%)	47.58	34.59	45.74	49.46

5.9 Summary

In this work, a novel approach for ABWE of narrowband speech is presented exploiting the sparse representation paradigm. The proposed ABWE approach is based on the premise that with a suitably learned over complete dictionary, it is more likely to produce similar sparse codes for NB and WB speech. Thus sparse codes for the NB speech obtained with NB dictionary when applied to WB dictionary can synthesize the bandwidth enhanced speech. The proposed approach is found to be quite effective for the voiced speech but for the unvoiced speech it required linear adaptation of sparse codes to enhance the performance. One possible extension of this work is to create the dictionaries and the linear transformation in joint manner rather than separately. Such an objective based learning of dictionary is already explored in form of semi-coupled dictionary in image processing domain.

A semi-coupled dictionary learning based ABWE is then developed for improving the SR-ABWE. Later a clustering approach is employed on the SCDL for further improvement. Among the different approaches proposed, the semi-coupled with clustering provides the best performance. The application of SR-ABWE method is demonstrated in children's ASR task using PFSTAR database. The SR indeed helps in reducing the mismatch between acoustic models trained using adults' speech data and tested using children's speech data.

6

Conclusions

Contents

The work started with an objective of developing methods for improving children's speech recognition under mismatched condition. The main distinction between children and adults is the change in the dimensions of speech production system and associated dynamics. Children's have shorter vocal tract length and hence important information about the shape of the vocal tract is present in the higher band (HB) ranging from 3.4-8 kHz. Also due to thinner dimensions of vocal folds, vibrations are faster compared to adults and hence the non-stationarity in case of children's speech is high. Due to both these aspects, the children's speech is quite different compared to that of adult. As a result, an ASR system trained using adults' speech gives poor performance when tested with children's speech.

Earlier attempts for improving children's speech recognition under mismatched condition included VTLN and minimizing pitch mismatch cases. In VTLN case, the vocal tract length is normalized so that the effect of shorter vocal tract length is minimized. The truncation of cepstral coefficients resulted in minimizing the pitch mismatch effect. The present work focused on a specific case termed narrowband testing scenario. However, the ASR is trained with WB adults' speech. When tested with both WB and NB adults' speech, the ASR performance is still good. Alternatively, testing with NB children's speech significantly degrades the performance to 9.37%. This is taken as mismatched condition in the present work and the question set for the exploration was how to improve ASR performance under mismatched condition. For this there should be some target performance value which can be set as the best achievable results. The ASR studies using WB speech was conducted. In case of adult, WB training and testing gives a WER of 0.35% and is 3.21 % for the case of WB children's speech. Therefore the best achievable performance that can be set as target for ASR under mismatched condition is 3.21%.

6.1 Summary

To carry out the exploration, the literature related to children's speech recognition was reviewed. Apart from other aspects, the two important directions explored based on VTLN and pitch mismatch cases were reviewed in detail. Both these approaches of course provide improved performance, but only by minimizing the effect of variabilities between children's

and adults' speech and not by adding or constructing any new information. However, what is required in the present work was to have methods that can reconstruct missing information in HB range for the children's speech. These methods are collectively termed as ABWE methods.

The first study started with the motivation to demonstrate the significance of ABWE for speech recognition. For this, a TI-DIGITS based digit recognition system using MFCC and HMM was developed. For comparison purpose, the ASR system was developed for both adults' and children's speech and also in each, NB and WB. When NB speech trained adults' speech models are tested with NB adults' speech, the performance is best demonstrating the matching condition. The NB children's speech tested against NB adults' speech models gave the worst performance indicating the significant mismatch. An existing mostly used ABWE method is implemented and ASR study is performed. The ABWE extended children's speech provided significant performance improvement compared to its NB case (4.06 %), demonstrating the significance of ABWE.

The VTLN using a standard approach alone also improved the ASR performance significantly in case of children's speech, for both NB (1.64%) and WB (0.77%) cases. This will tempt to conclude that VTLN will suffice to reduce mismatch in case of children's speech. However, the further improved performance in the case of combination of VTLN with ABWE (1.17%) as compared any of the individual cases infer that, there is some different information in both VTLN and ABWE methods. This is true also, as VTLN only performs normalization only within the given band, where as, ABWE tries to reconstruct information outside the band also.

The truncation of MFCC coefficients also provides significant performance improvement in children's speech (4.47%). This shows that the pitch and vocal tract length mismatch is significantly high in case of children's speech. Thus truncation of coefficients is effective in case of children's speech recognition under mismatched condition. Compared to NB, the WB case provides better performance for children's speech in the truncated features case (1.98%). This shows that, higher band information helps in children's speech recognition along with coefficients truncation. Therefore using ABWE along with coefficients truncation provided a performance which is significantly better compared to only by truncation of coefficients. Thus

the information offered by both the cases are different. The above studies strongly established the foundations for the proposed direction of work.

Having established the significance of ABWE for children's speech recognition under mismatched condition, the second part of the work concentrated on developing ABWE methods by exploiting some of the observations about the cause for the mismatch. These included class specific information, age specific information, and changes in vocal tract shape information in the form of delta features. To motivate the work initially, the statistical measures are described and computed. These included mutual information, differential entropy, their ratio and separability measure. The important observation that motivated the proposed method is the variability in the statistical measure values across class, age and delta features. Also, it was observed that the statistical measure values increase when the auxiliary information is exploited as compared to the global transform case.

The first method of ABWE was based on exploiting the class (digit) specific information. To see the effectiveness of the class specific information, initially supervised class specific ABWE method is proposed where the class information is taken from the available transcription. The global transform case gave 4.74 % and the supervised case gave 3.49 %. After an unsupervised class specific ABWE method is developed where the class information is predicted from the test data and used for ABWE and this provided 3.73 %. The results infer that the class specific information is indeed effective in performing ABWE.

A feature domain ABWE method is proposed using MFCC representation. The existing ABWE method based on MFCC follows the vocoder framework where the speech is synthesized from MFCC and then the ABWE is performed in vocoder framework. Alternatively, since the goal of present work is ASR, the ABWE using MFCC can be significantly simplified by working in the feature domain itself. Based on this motivation, a feature domain ABWE method using MFCC is proposed. The existing ABWE method based on MFCC provides a WER of 9.19 % where as the proposed one provides a WER of 8.88 %. This shows that even though improvement in terms of performance it is moderate, but the same is achieved with much simplified methodology.

The proposed feature domain MFCC based ABWE methods is then used for further demonstrating the significance of age specific and delta information. The NB performance for age specific study gave 9.83 %. The global transform case gave a WER of 9.0 %. The supervised case of age specific information provided a WER of 7.33 % which is the best performance that can be achieved using age information. The unsupervised age prediction method provides a WER of 7.31 % which is comparable to that of supervised case. The next study exploited the delta features information on top of the age specific information. The delta features information improves the performance of the system from 9.0 % to 6.23 % for global transform case and from 7.33 % to 5.15 % in case of age specific information case. Thus all these studies infer that it is indeed possible to develop ABWE methods using the auxiliary information and provide significantly better performance compared to the NB case.

Having exploited the auxiliary information for ABWE, the last part of the work focused on developing ABWE using the recent developments in the signal processing field. The sparse representation based signal modelling approach was found to result in state-of-the-art performances in many signal processing applications. Motivated by this, the present work also explored developing an ABWE method exploiting the sparse representation modelling of speech frames in the signal domain.

The proposed ABWE approach is based on the premise that with a suitably learned over complete dictionary, it is more likely to produce similar sparse codes for NB and WB speech. Thus sparse codes for the NB speech obtained with NB dictionary when applied to WB dictionary can synthesize the bandwidth enhanced speech. The proposed approach is found to be quite effective for the voiced speech but for the unvoiced speech it required linear adaptation of sparse codes to enhance the performance. A semi-coupled dictionary learning based ABWE is then developed for improving the SR-ABWE. Later a clustering approach is employed on the SCDL for further improvement. Among the different approaches proposed, the semi-coupled with clustering provided the best performance. The proposed SR-ABWE method gives better match between the acoustic models trained using adults' speech and tested using children's speech demonstrating the significance.

6.2 Major Contributions

The conclusions from the explorations made in this study are as follows:

- The ASR study using TI-DIGITS database infers that there is significant degradation in performance for children's speech recognition under mismatched condition.
- The VTLN approach improves the performance of children's speech recognition under mismatched condition by normalizing the vocal tract length. This effectively tries to minimize the mismatch due to difference in vocal tract length within the given band of speech.
- The coefficients truncation also improves the ASR performance by minimizing the pitch mismatch effect between children's and adults' speech.
- The use of standard ABWE is found to be effective for children's speech recognition under mismatched condition.
- The information offered by ABWE approach is different compared to those of VTLN and coefficients truncation, and can be used in combination with any of these methods for further improving performance.
- The ABWE using class-specific (digit-specific) information improves the ASR performance demonstrating its significance in ABWE.
- The proposed MFCC based ABWE method in the feature domain is effective like the existing MFCC based ABWE in vocoder framework, but with significant saving in computations as it operates in the feature domain.
- The age-specific information is also found to be useful for ABWE.
- The delta features representing the change in the vocal tract shape information indeed capture the fast changing dynamics associated with children's speech and hence the ABWE using this information is effective in improving ASR performance.

- The ABWE framework based on sparse representation seem to be effective in improving ASR performance.

- The semi-coupled dictionary based learning involving the clustering provides the best performance among the different sparse representation related explorations.

6.3 Future Work

- The exploration of sparse representation for ABWE is only an initial study. Further studies are needed to improve the performance. The same work can be explored for children's speech recognition under mismatched condition.

- The current work explored the use of GMM for mapping between NB and HB information. More recent methods like deep neural networks may be explored for capturing better relation/mutual information between NB and HB.

- The current work explored mostly MFCC based features for ABWE. The same ABWE methods can be explored using other features to see their effectiveness.

- Neural networks can be used for capturing the relations between NB and WB features. The same network can then be used for recovering WB features from NB features. The recovered features can be used for ASR studies.

- Methods may be developed for reducing the mismatch between children and adult speech by analyzing the excitation source information.

A

Objective Speech Quality Measures

To quantify the performance of the proposed ABWE approach, a number of measures have been computed. Among the different measures considered, the subband log spectral distortion (LSD) is the one which is designed to measure the performance of the bandwidth extension approaches. It measures the distortion in the missing frequency band and its square defined as

$$d_{LSD} = \sqrt{\frac{20}{\pi}\int_{w_l}^{w_h}\left(\log_{10}\frac{g}{|Y(e^{jw})|} - \log_{10}\frac{\hat{g}}{\left|\hat{Y}(e^{jw})\right|}\right)dw} \tag{A.1}$$

The segmental signal-to-noise ratio (segSNR) is a widely used objective measure. It requires the knowledge about both target and clean signals for computation and is defined as

$$\text{segSNR} = 10\log_{10}\left(\frac{\sum_{n=1}^{N} x^2[n]}{\sum_{n=1}^{N}(x[n]-\hat{x}[n])^2}\right) \tag{A.2}$$

where $x[n]$ is the clean speech, $\hat{x}[n]$ is the target speech, and N is the number of samples in the frame. Another popular measure is the log-likelihood ratio (LLR) which involves the LPC vectors computed from clean and target speech. LLR measure is calculated as

$$d_{LLR} = \log_{10}\left(\frac{\mathbf{a}_t\mathbf{R}_c\mathbf{a}_t^T}{\mathbf{a}_c\mathbf{R}_c\mathbf{a}_c^T}\right) \tag{A.3}$$

where $\mathbf{a}_c$ and $\mathbf{a}_t$ denote LPC vector for clean and target speech; $\mathbf{R}_c$ is the autocorrelation matrix for clean speech.

so that the test and reference patterns are compared with each other solely on the basis of their spectral shapes. The resulting distortion measure is called the likelihood ratio distortion measure and is represented as

$$d_{LR} = \frac{\mathbf{a}_t\mathbf{R}_c\mathbf{a}_t^T}{\mathbf{a}_c\mathbf{R}_c\mathbf{a}_c^T} - 1 \tag{A.4}$$

The Cepstrum Distance (CD) is an estimate of the log-spectrum distance between clean and distorted speech. Cepstrum is calculated by taking the logarithm of the spectrum and converting back to the time-domain. By going through this process, we can separate the speech excitation signal (pulse train signals from the glottis) from the convolved vocal tract characteristics. Cepstrum can also be calculated from LPC parameters with a recursion formula.

CD can be calculated as follows:

$$d_{CEP}(c_d, c_c) = \frac{10}{\log_{10}} \sqrt{2 \sum_{k=1}^{p} \{c_c(k) - c_d(k)\}} \tag{A.5}$$

where c_c and c_d are Cepstrum vectors for clean and distorted speech, and P is the order. Cepstrum distance is also a very efficient computation method of log-spectrum distance. It is more often used in speech recognition to match the input speech frame to the acoustic models.

The weighted likelihood ratio distortion measure, d_{WLR}, has the form

$$d_{WLR}^{N} = \sum_{i=1}^{N} \left(\frac{r_c(i)}{r_c(0)} - \frac{r_d(i)}{r_d(0)} \right) (c_c(k) - c_d(k)) \tag{A.6}$$

The weighted slope metric (WSM) has the form

$$d_{WSM} = \frac{1}{M} \sum_{m=1}^{M-1} \frac{\sum_{j=1}^{k} W(j,m)(S_c(j,m) - S_d(j,m))}{\sum_{j=1}^{k} W(j,m)} \tag{A.7}$$

where K is the number of bands, M is the total number of frames, and $S_c(j,m)$ and $S_d(j,m)$ are the spectral slopes (typically the spectral differences between neighboring bands) of the j^{th} band in the mth frame for clean and distorted speech respectively.

The perceptual evaluation of speech quality (PESQ) [24] is a quantitative measure that is designed to correlate with mean opinion score (MOS), a subjective measure of speech quality.

References

[1] S. Ghai, "Addressing pitch mismatch for children's automatic speech recognition," Ph.D. dissertation, Dept. of Electronics and Electrical Engg., Indian Institute of Technology Guwahati, India, Oct 2011.

[2] E. R. Larsen and R. M. Aarts, *Audio Bandwidth Extension: Application of Psychoacoustics, Signal Processing and Loudspeaker Design*. John Wiley and Sons, 2004.

[3] W. Krebber, "Voice transmission quality of telephone handsets (German)," Ph.D. dissertation, Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen, Germany, 1995.

[4] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *The Journal of the Acoustical Society of America*, vol. 19, no. 1, pp. 90–119, 1947.

[5] H. Pulakka, "Development and evaluation of artificial bandwidth extension methods for narrowband telephone speech," Ph.D. dissertation, Department of Signal Processing and Acoustics, School of Electrical Engineering, Aalto University, Aalto, Finlands, May 2013.

[6] M. Russell, S. D'Arcy, and L. Qun, "The effects of bandwidth reduction on human and computer recognition of children's speech," *Signal Processing Letters, IEEE*, vol. 14, no. 12, pp. 1044–1046, 2007.

[7] M. Russell, R. W. Series, J. L. Wallace, C. Brown, and A. Skilling, "The star system: an interactive pronunciation tutor for young children," *Computer Speech & Language*, vol. 14, no. 2, pp. 161–175, 2000.

[8] Q. Li and M. Russell, "An analysis of the causes of increased error rates in children's speech recognition," in *Proc.International Conference on Spoken Language Processing*, Denver,USA, 2002.

[9] M. Russell, "The PF-STAR British English Children's Speech Corpus," Dec 2006.

[10] M. Sanna and M. Murroni, "A codebook design method for fricative enhancement in artificial bandwidth extension," in *Proc. of the 5th International ICST Mobile Multimedia Communications Conference*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2009, p. 39.

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

[12] K.-Y. Park and H.-S. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," in *Proc. IEEE ICASSP*, vol. 3, 2000, pp. 1843–1846.

[13] P. Jax and P. Vary, "Wideband extension of telephone speech using a hidden markov model," in *Proc. IEEE Workshop on Speech Coding*, 2000, pp. 133–135.

[14] A. Uncini, F. Gobbi, and F. Piazza, "Frequency recovery of narrow-band speech using adaptive spline neural networks," in *Proc. IEEE ICASSP*, vol. 2, Mar 1999, pp. 997–1000.

[15] J. P. Cabral and L. C. Oliveira, "Pitch-synchronous time-scaling for prosodic and voice quality transformations," in *Proc. INTERSPEECH*, 2005, pp. 1137–1140.

[16] ——, "Pitch-synchronous time-scaling for highfrequency excitation regeneration," in *Proc. INTERSPEECH*, Sep 2005, pp. 1513–1516.

[17] 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 (EURASIP JASP)*, vol. 4, no. 4, pp. 266–274, Dec 2001.

[18] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[19] A. Houtsma, "Pitch and timbre: Definition, meaning and use," *Journal of New Music Research*, vol. 26, no. 2, pp. 104–115, 1997.

[20] U. Kornagel, "Spectral widening of the excitation signal for telephone-band speech enhancement," in *Proc. International Workshop on Acoustic, Echo and Noise Control (IWAENC)*, Darmstadt, Germany, Sep 2001, pp. 215–218.

[21] J. Epps and W. H. Holmes, "Speech enhancement using STC-based bandwidth extension," in *Proc. International Conference on Spoken Language Processing*, vol. 2, Sydney, Australia, Nov 1998, pp. 519–522.

[22] B. Iser, W. Minker, and G. Schmidt, *Bandwidth Extension of Speech Signals*, 1st ed. Springer Publishing Company, Inc., 2008.

[23] S. A. V. Nels Rohde, "Artificial bandwidth extension of narrowband speech," Master's thesis, Department of Electronic Systems,Aalborg University, Aalborg, Jun 2007.

[24] ITU-T P.862.2, "ITU-T recommendation P.862.2: Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," International Telecommunication Union, Tech. Rep., 2007.

[25] P. Vary and B. Geiser, "Steganographic wideband telephony using narrowband speech codecs," in *Proc. Forty-First Asilomar Conference on Signals, Systems and Computers (ACSSC)*, Nov 2007, pp. 1475–1479.

[26] G. Fuchs and R. Lefebvre, "A new post-filtering for artificially replicated high-band in speech coders," in *Proc. IEEE ICASSP*, vol. 1, May 2006, pp. 1–4.

[27] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," in *Proc. IEEE ICASSP*, vol. 4, Apr 1979, pp. 428–431.

[28] J. M. Valin and R. Lefebvre, "Bandwidth extension of narrowband speech for low bit-rate wideband coding," in *Proc IEEE Workshop on Speech Coding*, 2000, pp. 130–132.

[29] J. S. Park, M. Y. Choi, and H. S. Kim, "Low-band extension of celp speech coder by harmonics recovery," in *Proc. of International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, Nov 2004, pp. 147–150.

[30] A. McCree, "A 14 kb/s wideband speech coder with a parametric highband model," in *Proc. IEEE ICASSP*, vol. 2, 2000, pp. 1153–1156.

[31] L. Laaksonen, J. Kontio, and P. Alku, "Artificial bandwidth expansion method to improve intelligibility and quality of AMR-coded narrowband speech," in *Proc. IEEE ICASSP*, vol. 1, Mar 2005, pp. 809–812.

[32] I. Varga, S. Proust, and H. Taddei, "ITU-T G.729.1 scalable codec for new wideband services," in *IEEE Communications Magazine*, vol. 47, no. 10, Oct 2009, pp. 131–137.

[33] C. Avendano, H. Hermansky, and E. A. Wan, "Beyond nyquist: Towards the recovery of broad-bandwidth speech from narrow-bandwidth speech," in *Proc. EUROSPEECH*, 1995, pp. 165–168.

[34] V. Berisha and A. Spanias, "Wideband speech recovery using psychoacoustic criteria," *EURASIP Journal on Audio, Speech, and Music Processing (EURASIP JASMP)*, vol. 2, no. 2, pp. 1–5, Apr 2007.

[35] B. C. J. Moore and C.-T. Tan, "Perceived naturalness of spectrally distorted speech and music," *The Journal of the Acoustical Society of America*, vol. 114, no. 1, pp. 408–419, 2003.

[36] S. Voran, "Listener ratings of speech passbands," in *Proc. IEEE Workshop on Speech Coding For Telecommunications*, Sep 1997, pp. 81–82.

[37] S. Steidl, G. Stemmer, C. Hacker, E. Nöth, and H. Niemann, "Improving children's speech recognition by HMM interpolation with an adults' speech recognizer," in *Proc. annual pattern recognition symposium of the German Association for Pattern Recognition (DAGM)*, 2003, pp. 600–607.

[38] D. Giuliani and M. Gerosa, "Investigating recognition of children's speech," in *Proc. IEEE ICASSP*, vol. 2, 2003, pp. 137–140.

[39] B. L. Smith, S. H. Long, and M. D. Sugarman, "Experimental manipulation of speaking rate for studying temporal variability in children's speech," *The Journal of the Acoustical Society of America*, vol. 72, no. S1, pp. S64–S64, 1982.

[40] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, Nov 2003.

[41] C. A. Bickley, "Acoustic evidence for the development of speech," Ph.D. dissertation, Research Laboratory of Electronics, Massachusetts Institute of Technology, Oct 1989.

[42] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[43] A. Potamianos and S. Narayanan, "A review of the acoustic and linguistic properties of children's speech," in *Proc. 9th Workshop on Multimedia Signal Processing, IEEE*, 2007, pp. 22–25.

[44] S. Schötz, "A perceptual study of speaker age," in *Working paper 49, Lund University, Dept of Linguistic*, 2001, pp. 136–139.

[45] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 2, pp. 65–78, Feb 2002.

[46] E. F. Strommen and F. S. Frome, "Talking back to big bird: Preschool users and a simple speech recognition system," in *Proc. Educational Technology Research and Development*, vol. 41, no. 1, Feb 1993, pp. 5–16.

[47] A. Potamianos and S. Narayanan, "Spoken dialog systems for children," in *Proc. IEEE ICASSP*, vol. 1, May 1998, pp. 197–200.

[48] V. Farantouri, A. Potamianos, and S. Narayanan, "Linguistic analysis of spontaneous children's speech," in *Proc. Workshop on Child, Computer and Interaction, Chania, Crete, Greece*, Oct 2008.

[49] V. Viswanathan and W. Lai, "Synthesis of high-pitched sounds," Jul. 9 2002, US Patent 6,418,406. [Online]. Available: https://www.google.com.ar/patents/US6418406

[50] F. Villavicencio, A. Robel, and X. Rodet, "Improving LPC spectral envelope extraction of voiced speech by true-envelope estimation," in *Proc. IEEE ICASSP*, vol. 1, May 2006, pp. 1–4.

[51] S. L. Nissen and R. A. Fox, "Acoustic and spectral characteristics of young children's fricative productions: A developmental perspectives," *The Journal of the Acoustical Society of America*, vol. 118, no. 4, pp. 2570–2578, 2005.

[52] ——, "Acoustic and spectral patterns in young children's stop consonant productions," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1369–1378, 2009.

[53] R. Sinha and S. Ghai, "On the use of pitch normalization for improving children's speech recognition," in *Proc. INTERSPEECH*, Brighton, UK, Sep 2009, pp. 568–571.

[54] S. Ghai and R. Sinha, "Exploring the role of spectral smoothing in context of children's speech recognition," in *Proc. INTERSPEECH*, 2009, pp. 1607–1610.

[55] ——, "Exploring the effect of differences in the acoustic correlates of adults' and children's speech in the context of automatic speech recognition," *EURASIP J. Audio Speech Music Process.*, pp. 7:1–7:15, Jan 2010.

[56] D. Burnett and M. Fanty, "Rapid unsupervised adaptation to children's speech on a connected-digit task," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, vol. 2, Oct 1996, pp. 1145–1148.

[57] P. Cosi and B. L. Pellom, "Italian children's speech recognition for advanced interactive literacy tutors," in *Proc. INTERSPEECH*, Lisbon, 2005, pp. 2201–2204.

[58] D. Elenius and M. Blomberg, "Adaptation and normalization experiments in speech recognition for 4 to 8 year old children," in *Proc. INTERSPEECH*, 2005, pp. 2749–2752.

[59] D. Giuliani and M. Gerosa, "Investigating recognition of children's speech," in *Proc. IEEE ICASSP*, vol. 2, Apr 2003, pp. 137–140.

[60] J. Gustafson and K. Sjölander, "Voice transformations for improving children's speech recognition in a publicly available dialogue system," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, Denver, 2002, pp. 297–300.

[61] A. Hagen, "Advances in children's speech recognition with application to interactive literacy tutors," Ph.D. dissertation, Boulder, CO, USA, 2006.

[62] S. Molau, S. Kanthak, and H. Ney, "Efficient vocal tract normalization in automatic speech recognition," in *Proc. Electronic Speech Signal Processing (ESSV)*, Cottbus, Germany, 2000, pp. 209–216.

[63] A. Potamianos and R. Rose, "On combining frequency warping and spectral shaping in HMM based speech recognition," in *Proc. IEEE ICASSP*, Munich, 1997, pp. 1275–1278.

[64] T. Claes, I. Dologlou, L. ten Bosch, and D. Van Compernolle, "A novel feature transformation for vocal tract length normalization in automatic speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 6, pp. 549–557, Nov 1998.

[65] S. Das, D. Nix, and M. Picheny, "Improvements in children's speech recognition performance," in *Proc. IEEE ICASSP*, vol. 1, May 1998, pp. 433–436.

[66] G. Stemmer, C. Hacker, S. Steidl, and E. Nöth, "Acoustic normalization of children's speech," in *Proc. INTERSPEECH*, 2003, pp. 1313–1316.

[67] M. Gerosa, "Acoustic modeling for automatic recognition of children's speech," Ph.D. dissertation, University of Trento, 2006.

[68] S. D'Arcy, L. Wong, and M. Russell, "Recognition of read and spontaneous children's speech using two new corpora," in *Proc. INTERSPEECH*, Jeju Island, Korea, 2004, pp. 1473–1476.

[69] S. Steidl, G. Stemmer, C. Hacker, E. Nöth, and H. Niemann, "Improving children's speech recognition by HMM interpolation with an adults' speech recognizer," in *Pattern Recognition*, ser. Lecture Notes in Computer Science, B. Michaelis and G. Krell, Eds. Springer Berlin Heidelberg, 2003, vol. 2781, pp. 600–607.

[70] A. Andreaou, T. Kamm, and J. Cohen, "Experiments in vocal tract length normalization," in *Proc. Computer Analysis of Images and Patterns (CAIP) Workshop: Frontiers in Speech Recognition*, 1994.

[71] L. Lee and R. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. IEEE ICASSP*, vol. 1, May 1996, pp. 353–356.

[72] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*. Cambridge University Press, 2006.

[73] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 1, pp. 49–60, Jan 1998.

[74] J. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, Apr 1994.

[75] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[76] "Paired comparison test of wideband and narrowband telephony," International Telecommunication Union, Tech. Rep. COM 12-9-E, Mar 1993.

[77] A. Rämö and H. Toukomaa, "On comparing speech quality of various narrow-and wideband speech codecs," in *Proc. of the Eighth International Symposium on in Signal Processing and Its Applications (ISSPA)*, 2005, pp. 603–606.

[78] S. Möller, M. Wältermann, B. Lewcio, N. Kirschnick, and P. Vidales, "Speech quality while roaming in next generation networks," in *Proc. IEEE International Conference on Communications (ICC)*, 2009, pp. 1–5.

[79] S. Voran, "Subjective ratings of instantaneous and gradual transitions from narrowband to wideband active speech," in *Proc. IEEE ICASSP*, Mar 2010, pp. 4674–4677.

[80] P. Jax and P. Vary, "Artificial bandwidth extension of speech signals using MMSE estimation based on a hidden markov model," in *Proc. IEEE ICASSP*, vol. 1, 2003, pp. 680–683.

[81] Y. Qian and P. Kabal, "Combining equalization and estimation for bandwidth extension of narrowband speech," in *Proc. IEEE ICASSP*, vol. 1, May 2004, pp. 713–716.

[82] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding And Error Concealment*. John Wiley and Sons, 2006.

[83] C. Beaugeant, M. Schönle, and I. Varga, "Challenges of 16 khz in acoustic pre-and post-processing for terminals," *IEEE Communications Magazine*, vol. 44, no. 5, pp. 98–104, 2006.

[84] R. V. Cox, S. F. De Campos Neto, C. Lamblin, and M. H. Sherif, "ITU-T coders for wideband, superwideband, and fullband speech communication [series editorial]," *IEEE Communications Magazine*, vol. 47, no. 10, pp. 106–109, 2009.

[85] Y. Agiomyrgiannakis and Y. Stylianou, "Conditional vector quantization for speech coding," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 377–386, 2007.

[86] A. H. Nour-Eldin and P. Kabal, "Memory-based approximation of the gaussian mixture model framework for bandwidth extension of narrowband speech." in *Proc. INTERSPEECH*, 2011, pp. 1185–1188.

[87] P. Jax and P. Vary, "An upper bound on the quality of artificial bandwidth extension of narrowband speech signals," in *Proc. IEEE ICASSP*, vol. 1, 2002, pp. 237–240.

[88] M. Nilsson, H. Gustaftson, S. V. Andersen, and W. B. Kleijn, "Gaussian mixture model based mutual information estimation between frequency bands in speech," in *Proc. IEEE ICASSP*, vol. 1, 2002, pp. 525–528.

[89] A. H. Nour-Eldin and P. Kabal, "Objective analysis of the effect of memory inclusion on bandwidth extension of narrowband speech," in *Proc. INTERSPEECH*, 2007, pp. 2489–2492.

[90] A. Nour-Eldin and P. Kabal, "Combining frontend-based memory with MFCC features for bandwidth extension of narrowband speech," in *Proc. IEEE ICASSP*, 2009, pp. 4001–4004.

[91] U. Kornagel, "Techniques for artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 86, no. 6, pp. 1296–1306, 2006.

[92] K.-T. Kim, M.-K. Lee, and H.-G. Kang, "Speech bandwidth extension using temporal envelope modeling," *Signal Processing Letters, IEEE*, vol. 15, pp. 429–432, 2008.

[93] J. Epps, "Wideband extension of narrowband speech for enhancement and coding," Ph.D. dissertation, The University of New South Wales, Australia, Sep 2000.

[94] Y. Agiomyrgiannakis and Y. Stylianou, "Combined estimation/coding of highband spectral envelopes for speech spectrum expansion," in *Proc. IEEE ICASSP*, vol. 1, 2004, pp. 469–472.

[95] B. Geiser, P. Jax, P. Vary, H. Taddei, S. Schandl, M. Gartner, C. Guillaume, and S. Ragot, "Bandwidth Extension for Hierarchical Speech and Audio Coding in ITU-T Rec. G.729.1," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2496–2509, Nov 2007.

[96] A. Nishimura, "Steganographic band width extension for the AMR codec of low-bit-rate modes," in *Proc. INTERSPEECH*, 2009, pp. 2611–2614.

[97] B. Geiser and P. Vary, "Backwards compatible wideband telephony in mobile networks: CELP watermarking and bandwidth extension," in *Proc. IEEE ICASSP*, vol. 4, Apr 2007, pp. 533–536.

[98] A. Sagi and D. Malah, "Bandwidth extension of telephone speech aided by data embedding," *EURASIP Journal on Applied Signal Processing (EURASIP JASP)*, vol. 1, no. 1, pp. 37–37, 2007.

[99] M. G. Croll, "Sound-quality improvement of broadcast telephone calls," The British Broadcasting Corporation (BBC), Research Department, Tech. Rep. 26, 1972.

[100] P. Jax, "Enhancement of bandlimited speech signals: Algorithms and theoretical bounds," Ph.D. dissertation, Chair and Institute of Communication Systems and Data Processing, RWTH, 2002.

[101] Y. M. Cheng, D. O'Shaughnessy, and P. Mermelstein, "Statistical recovery of wideband speech from narrowband speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 544–548, Oct 1994.

[102] H. Carl and U. Heute, "Bandwidth enhancement of narrow-band speech signals," in *Proc. EUSIPCO*, 1994, pp. 1178–1181.

[103] Y. Yoshida and M. Abe, "An algorithm to reconstruct wideband speech from narrowband speech based on codebook mapping," in *Proc. International Conference on Spoken Language Processing (ICSLP), Yokohama, Japan*, Sep 1994, pp. 1591–1594.

[104] M. Nilsson and W. Kleijn, "Avoiding over-estimation in bandwidth extension of telephony speech," in *Proc. IEEE ICASSP*, vol. 2, 2001, pp. 869–872.

[105] I. Katsir, I. Cohen, and D. Malah, "Speech bandwidth extension based on speech phonetic content and speaker vocal tract shape estimation," *Proc. EUSIPCO*, pp. 461–465, 2011.

[106] J. Epps and W. H. Holmes, "Speech enhancement using STC-based bandwidth extension," in *Proc. International Conference on Spoken Language (ICSLP)*, 1998.

[107] G. Miet, A. Gerrits, and J. Valiere, "Low-band extension of telephone-band speech," in *Proc. IEEE ICASSP*, vol. 3, 2000, pp. 1851–1854.

[108] B. Iser and G. Schmidt, "Neural networks versus codebooks in an application for bandwidth extension of speech signals," in *Proc. EUROSPEECH*, Geneva, Switzerland, Sep 2003, pp. 565–568.

[109] T. Pham, F. Schaefer, and G. Kubin, "A novel implementation of the spectral shaping approach for artificial bandwidth extension," in *Proc. Third International Conference on Communications and Electronics (ICCE)*, Aug 2010, pp. 262–267.

[110] Y. Qian and P. Kabal, "Dual-mode wideband speech recovery from narrowband speech." in *Proc. INTERSPEECH*, 2003.

[111] R. Hu, V. Krishnan, and D. V. Anderson, "Speech bandwidth extension by improved codebook mapping towards increased phonetic classification." in *Proc. INTERSPEECH*, 2005, pp. 1501–1504.

[112] T. Unno and A. McCree, "A robust narrowband to wideband extension system featuring enhanced codebook mapping," in *Proc. IEEE ICASSP*, vol. 1, Mar 2005, pp. 805–808.

[113] T. Ramabadran and M. Jasiuk, "Artificial bandwidth extension of narrow-band speech signals via high-band energy estimation," in *Proc. EUSIPCO*, 2008, pp. 1–5.

[114] B. Geiser and P. Vary, "Beyond wideband telephony - bandwidth extension for super-wideband speech." in *Proc. of German Annual Conference on Acoustics (DAGA)*, 2008, pp. 635–636.

[115] M. R. P. Thomas, J. Gudnason, P. Naylor, B. Geiser, and P. Vary, "Voice source estimation for artificial bandwidth extension of telephone speech," in *Proc. IEEE ICASSP*, Mar 2010, pp. 4794–4797.

[116] ITU-T G.729.1, "ITU-T recommendation G.729.1: G.729 based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729," International Telecommunication Union, Tech. Rep., 2006.

[117] P. Jax and P. Vary, "Feature selection for improved bandwidth extension of speech signals," in *Proc. IEEE ICASSP*, vol. 1, 2004, pp. 697–700.

[118] J. Kontio, L. Laaksonen, and P. Alku, "Neural network-based artificial bandwidth expansion of speech," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 873–881, Mar 2007.

[119] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter, "Speech enhancement via frequency bandwidth extension using line spectral frequencies," in *Proc. IEEE ICASSP*, vol. 1, 2001, pp. 665–668.

[120] S. Vaseghi, E. Zavarehei, and Q. Yan, "Speech bandwidth extension: extrapolations of spectral envelop and harmonicity quality of excitation," in *Proc. IEEE ICASSP*, vol. 3, 2006, pp. 1–4.

[121] S. Yao and C.-F. Chan, "Speech bandwidth enhancement using state space speech dynamics," in *Proc. IEEE ICASSP*, vol. 1, 2006, pp. 1–4.

[122] C. Yagh and E. Erzin, "Artificial bandwidth extension of spectral envelope with temporal clustering," in *Proc. IEEE ICASSP*, 2011, pp. 5096–5099.

[123] C. Liu, Q.-J. Fu, and S. S. Narayanan, "Effect of bandwidth extension to telephone speech recognition in cochlear implant users," *The Journal of the Acoustical Society of America*, vol. 125, no. 2, pp. 77–83, 2009.

[124] A. Shahina and B. Yegnanarayana, "Mapping neural networks for bandwidth extension of narrowband speech," in *Proc. INTERSPEECH*, Sep 2006, pp. 1435–1438.

[125] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[126] D. O'Shaughnessy, "Invited paper: Automatic speech recognition: History, methods and challenges," *Pattern Recognition*, vol. 41, no. 10, pp. 2965–2979, 2008.

[127] M. L. Seltzer and A. Acero, "Training wideband acoustic models using mixed-bandwidth training data for speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 235–245, 2007.

[128] ——, "Training wideband acoustic models using mixed-bandwidth training data via feature bandwidth extension." in *Proc. IEEE ICASSP*, vol. 1, 2005, pp. 921–924.

[129] G.-B. Song and P. Martynovich, "A study of HMM-based bandwidth extension of speech signals," *Signal Processing*, vol. 89, no. 10, pp. 2036–2044, 2009.

[130] P. Dymarski, N. Moreau, and G. Richard, "Greedy sparse decompositions: a comparative study," *EURASIP Journal on Advances in Signal Processing*, vol. 34, 2011.

[131] J. Yang, Y. Peng, W. Xu, and Q. Dai, "Ways to sparse representation: A comparative study," *Tsinghua Science and Technology*, vol. 14, no. 4, pp. 434–443, Aug 2009.

[132] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov 2006.

[133] S. Wang, D. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, Jun 2012, pp. 2216–2223.

[134] R. Leonard, "A database for speaker-independent digit recognition," in *Proc. IEEE ICASSP*, vol. 9, Mar 1984, pp. 328–331.

[135] J. E. Markel and A. H. Gray, *Linear Prediction of Speech*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1982.

[136] L. Ljung, Ed., *System Identification (2nd ed.): Theory for the User*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1999.

[137] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, united states ed. Prentice Hall, Apr 1993.

[138] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Waltham, USA: Academic Press, 1990.

[139] A. H. Nour-Eldin and P. Kabal, "Mel-frequency cepstral coefficient-based bandwidth extension of narrowband speech." in *Proc. INTERSPEECH*, 2008, pp. 53–56.

[140] Y. Sunil and R. Sinha, "Exploration of class specific ABWE for robust children's ASR under mismatched condition," in *Proc. International Conference on Signal Processing and Communications (SPCOM)*, Jul 2012, pp. 1–5.

[141] Y. Sunil, S. Ghai, and R. Sinha, "Exploration of artificial bandwidth expansion for improving children's ASR in mismatched condition," in *Proc. Centenary Conference*, Electrical Engineering, Indian Institute of Science, Bangalore, India, 2011, pp. 143–147.

[142] M. Nilsson, H. Gustaftson, S. Andersen, and W. Kleijn, "Gaussian mixture model based mutual information estimation between frequency bands in speech," in *Proc. IEEE ICASSP*, vol. 1, 2002, pp. 525–528.

[143] A. Nour-Eldin, T. Shabestary, and P. Kabal, "The effect of memory inclusion on mutual information between speech frequency bands," in *Proc. IEEE ICASSP*, vol. 3, May 2006, pp. 1–4.

[144] B. Pellom and J. Hansen, "Voice analysis in adverse conditions: the centennial olympic park bombing 911 call," in *Proc. of the 40th Midwest Symposium on Circuits and Systems (MWSCAS)*, vol. 2, Aug 1997, pp. 873–876.

[145] I. Mporas and T. Ganchev, "Estimation of unknown speaker's height from speech," *International Journal of Speech Technology*, vol. 12, no. 4, pp. 149–160, 2009.

[146] Y. Qian and P. Kabal, "Combining equalization and estimation for bandwidth extension of narrowband speech," in *Proc. IEEE ICASSP*, vol. 1, May 2004, pp. 713–716.

[147] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, Dec 2006.

[148] L. Fang, S. Li, X. Kang, J. A. Izatt, and S. Farsiu, "3-d adaptive sparsity based image compression with applications to optical coherence tomography," *IEEE Transactions on Medical Imaging*, vol. 34, no. 6, pp. 1306–1320, June 2015.

[149] B. C. Haris and R. Sinha, "Robust speaker verification with joint sparse coding over learned dictionaries," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 10, pp. 2143–2157, Oct 2015.

[150] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, Sept 2011.

[151] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, Nov 2010.

[152] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition," in *Proc. IEEE ICASSP*, vol. 1, Detroit, 1995, pp. 81–84.

[153] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. New York, NY, USA: Elsevier Science Inc., 1995.

[154] S. Ghai and R. Sinha, "Pitch adaptive MFCC features for improving children's mismatched ASR," *International Journal of Speech Technology*, vol. 18, no. 3, pp. 489–503, 2015.

List of Publications

- Published Paper and Accepted Publication:

  1. Sunil Y, Rohit Sinha, "Sparse Representation Based Approach for Artificial Bandwidth Extension of Speech," in Proc. IEEE SPCOM, pp. 1-5, 2014.
  2. Sunil Y, Rohit Sinha, "Exploration of MFCC based ABWE for robust Children's speech recognition under mismatched condition," in Proc. IEEE SPCOM, pp. 1-5, 2014.
  3. Sunil Y, Rohit Sinha, "Exploration of class specific ABWE for robust Children's ASR under mismatched condition," in Proc. IEEE SPCOM, pp. 1-5, 2012.
  4. Sunil Y, S. Ghai and Rohit Sinha, "Exploration of artificial bandwidth expansion for improving children's ASR in mismatched condition ", in proc. Centenary Conference,Electrical Engineering, Indian Institute of Science, pp. 143–147, 2011.

- Manuscripts accepted for publication

  1. Sunil Y, S. R. Mahadeva Prasanna and Rohit Sinha, "Children's Speech Recognition Under Mismatched Condition: A Review," [accepted for publication] IETE Journal of Education, 2016.

- Manuscripts to be communicated

  1. Sunil Y, S. R. Mahadeva Prasanna and Rohit Sinha, "Combining approaches for improving children's speech recognition under mismatched condition," [to be communicated].
  2. Ganji Sreeram, Sunil Y, Rohit Sinha and S. R. Mahadeva Prasanna, "Improvements in Sparse Representation based Automatic Bandwidth Extension of Speech," [to be communicated].