ADDRESSING PITCH MISMATCH FOR CHILDREN'S AUTOMATIC SPEECH RECOGNITION

A

thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

by

SHWETA GHAI

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781039, INDIA

OCTOBER 2011

Dedicated to,

My Parents

Sh. R. K. Ghai and Smt. Promila Ghai

for their love, support and blessings

My Husband

Gaurav Bhatia

for his patience, cooperation and understanding

AND

The Almighty

Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati
Guwahati - 781039, India.

Certificate

*This is to certify that the thesis entitled "***ADDRESSING PITCH MISMATCH FOR CHILDREN'S AUTOMATIC SPEECH RECOGNITION***", submitted by* ***Shweta Ghai****, Roll No. 06610209, 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 her under my supervision and guidance. The work contained in this thesis has not been submitted elsewhere for the award of any degree.*

Dated:

Dr. Rohit Sinha
Associate Professor

Acknowledgements

I express my deepest and most sincere gratitude to my supervisor **Dr. Rohit Sinha** for his guidance and constant encouragement. His insightful feedbacks have helped me greatly in improving the quality of my thesis. I greatly admire his attitude towards research, creative thinking and enthusiasm for work.

I am also very thankful to my doctoral committee members **Prof. S. Dandapat**, **Prof. P. K. Bora**, **Dr. S. R. M. Prasanna** and **Dr. P. K. Das** for their valuable suggestions and for sparing their precious time to evaluate the progress of my work.

I owe my invaluable thanks to Dr. João P. Cabral, visiting researcher at IITG who helped me during the initial stages of my research work in developing some programmes. I sincerely thank Mr. Shakti Prasad Rath, research scholar, IITM for helping me in getting through various bottlenecks during my research work. I also thank Dr. P. Krishnamoorthy, Dr. M. Sabarimalai Manikandan and Dr. H. S. Jayanna for their kind help.

I am grateful to all the technical staff members of the department without whose help I could not have completed my experiments. My special thanks to Mr. L. N. Sharma, Mr. Sanjib Das and Mrs. D. Jharna for maintaining the computing facility and various resources useful for the research work to their best and for their timely help.

I sincerely thank Mrs. Nirmala for her help and care rendered to me whenever needed. I thank all my dear friends Geetika, Maya, Mamta and Sumitra for the love, care and help given to me during my stay at IITG. Thanks go out to all my friends at the Electro Medical and Speech Technology Laboratory for creating a healthy research environment. Special thanks also go to Debadatta Pati and Haris for their help during my stay.

My deepest gratitude goes to my parents for their continuous love and support throughout my career. The opportunities that they have given me and their unlimited sacrifices are the reasons where I am and what I have accomplished so far.

Shweta Ghai

Abstract

This thesis addresses the acoustic mismatch due to pitch differences between the adults' and the children's speech for children's automatic speech recognition (ASR) on adults' speech trained models. The motivation for the work is obtained through the study done on exploring various acoustic sources of mismatch: pitch, speaking rate, formant frequencies and glottal flow parameters (open quotient, return quotient and speed quotient) for children's ASR on adults' speech trained models. The effect of variations in each of these acoustic correlates across speech signals is studied on Mel frequency cepstral coefficient (MFCC) features and ASR models. Following that, their relative significance is explored for children's speech recognition on the adults' speech trained models in a consistent setup. It is found that apart from the formant frequencies, the pitch is the other major source of acoustic mismatch between the adults' and the 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. Motivated by that, the pitch-robustness of perceptual linear prediction cepstral coefficient (PLPCC) and perceptual minimum variance distortionless response (PMVDR) cepstral coefficient features is studied to explore their efficacy for children's ASR on adults' speech trained models in comparison to MFCC features. It is found that MFCC features outperform PLPCC features while with suitable optimization of model order PMVDR features are more pitch-robust than MFCC features. However, the children's ASR performance obtained with MFCC features after explicit pitch normalization of children's speech is found to be comparable to that obtained with PMVDR features after optimization of its model order for children's speech. Following the observations, 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 be-

tween 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. Using the proposed adaptive MFCC feature truncation algorithm significant improvements are found in the children's and adults' ASR performances on children's speech trained models as well. The improvements obtained in the ASR performances with the proposed algorithms are also found to be additive to those obtained with the existing speaker normalization and model adaptation techniques viz., VTLN, MLLR and CMLLR.

Keywords: Children's speech recognition, acoustic mismatch, pitch, speaking rate, glottal flow parameters, MFCC, PLPCC, PMVDR, Mel filterbank, cepstral truncation.

Contents

List of Figures xv

List of Tables xxi

List of Acronyms xxvii

1 Introduction 1

1.1 Overview of Automatic Speech Recognition . . . 2

1.2 Challenges in Children's Speech Recognition . . . 3

1.2.1 Acoustic Correlates of Children's Speech . . . 4

1.2.2 Linguistic Correlates of Children's Speech . . . 5

1.3 Performances of ASR Systems for Children's Speech . . . 6

1.4 Review of Approaches used for Children's ASR . . . 7

1.4.1 Acoustic Mismatch . . . 7

1.4.1.1 Feature Domain Approaches . . . 7

1.4.1.2 Signal Domain Approaches . . . 9

1.4.1.3 Model Domain Approaches . . . 10

1.4.2 Linguistic Mismatch . . . 12

1.5 Motivation of the Thesis . . . 13

1.6 Objectives of the Thesis . . . 14

1.7 Organization of the Thesis . . . 14

2 Speech Corpora and Experimental Setups 17

2.1 Introduction . . . 18

2.2 Speech Corpora . . . 18

2.3 Speech Recognition Systems . . . 19

2.3.1 Connected Digit Recognition . . . 21

2.3.2 Continuous Speech Recognition . . . 21
2.4 Methods for Transformation of Acoustic Correlates of Speech . . . 22
2.4.1 Signal Domain Method: PSTS . . . 23
2.4.1.1 Transformation of Pitch and Signal Duration . . . 24
2.4.1.2 Transformation of Glottal Flow Parameters . . . 25
2.4.2 Feature Domain Method: VTLN . . . 27
2.5 Model Adaptation Techniques . . . 28
2.5.1 MLLR . . . 29
2.5.1.1 MLLR-MEAN . . . 29
2.5.1.2 MLLR-COV . . . 30
2.5.2 CMLLR . . . 30
2.6 Summary . . . 32
3 Role of Various Acoustic Sources of Mismatch in Children's ASR 33
3.1 Introduction . . . 34
3.2 Effect of Various Acoustic Sources of Mismatch on MFCC Features & ASR Models . . . 35
3.2.1 Pitch . . . 35
3.2.2 Speaking Rate . . . 40
3.2.3 Glottal Flow Parameters . . . 42
3.2.4 Formant Frequencies . . . 43
3.3 Relative Significance of Various Acoustic Sources of Mismatch for Children's ASR . . . 46
3.3.1 Connected Digit Recognition Task . . . 47
3.3.1.1 Pitch . . . 47
3.3.1.2 Speaking Rate . . . 50
3.3.1.3 Glottal Flow Parameters . . . 52
3.3.1.4 Formant Frequencies . . . 53
3.3.2 Continuous Speech Recognition Task . . . 53
3.4 Combining VTLN and Explicit Acoustic Normalization with Model Adaptation . . . 56
3.5 Summary . . . 58
4 Effect of Pitch on MFCC Features 61
4.1 Introduction . . . 62

4.2 Effect of Uniform and Non-Uniform Filterbank on Pitch Harmonicity 63
4.2.1 Uniform Filterbank based Spectral Analysis 63
4.2.2 Non-Uniform Filterbank based Spectral Analysis 66
4.3 Effect of Pitch-dependent Distortions on MFCCs 71
4.4 Summary 74

5 Pitch-Robustness of Salient ASR Features for Children's ASR 77

5.1 Introduction 78
5.2 Efficacy of PLPCC Features for Children's ASR 79
5.2.1 Effect of Pitch on PLPCC Features 80
5.2.2 Children's Speech Recognition using PLPCC Features 84
5.3 Children's ASR using PMVDR Features 86
5.4 Summary 90

6 Pitch Normalization by Filterbank Modification 93

6.1 Introduction 94
6.2 Mel Filterbank Modification for Pitch Normalization 95
6.2.1 Implicit Modification of Filter Bandwidths 95
6.2.2 Selective Modification of Filter Bandwidths 99
6.3 Proposed Pitch Normalization Algorithm 105
6.4 Combining Proposed Algorithm with VTLN and CMLLR 107
6.5 Summary 109

7 Pitch Mismatch Reduction by Cepstral Truncation 111

7.1 Introduction 112
7.2 Truncation of MFCC Features for Children's ASR 112
7.3 Role of MFCC Feature Truncation in Pitch Mismatch Reduction 118
7.4 Adaptive MFCC Feature Truncation for Pitch Mismatch Reduction 122
7.4.1 Correlation between MFCC Feature Truncation and VTLN Warp Factor 122
7.4.2 Proposed Algorithm for Adults' Speech Trained ASR Models 123
7.4.3 Proposed Algorithm for Children's Speech Trained ASR Models 129
7.5 Combining Proposed Algorithm with VTLN and CMLLR 133
7.6 Summary 135

8 Summary and Future Work **137**

8.1 Summary of the Work 138

8.2 Contributions of the Work 140

8.3 Scope for the Future Work 141

A Mel Frequency Cepstral Coefficients **143**

A.1 Mel Frequency Cepstral Coefficient (MFCC) Computation 144

B Hidden Markov Models **147**

B.1 Hidden Markov Model (HMM) 148

B.1.1 Training 148

B.1.2 Testing 151

C Perceptual Linear Prediction Cepstral Coefficients **153**

C.1 Perceptual Linear Prediction Cepstral Coefficient (PLPCC) Computation 154

D Perceptual-MVDR Cepstral Coefficients **159**

D.1 Perceptual Minimum Variance Distortionless Response (PMVDR) Cepstral Coefficient Computation 160

Bibliography **163**

List of Publications **171**

List of Figures

2.1 Representation of the extracted time instants and the glottal cycle phases in (a) the glottal flow waveform and (b) its time-derivative (i.e., the LP residual signal). The figure is adapted from [1]. 26

3.1 Plots of the signals and the smoothed Mel spectra (referred to as 'Smoothed') along with their corresponding linear DFT spectra for central steady-state portions of vowel /IY/ having pitch values of around (a) 100 Hz (b) 220 Hz (c) 300 Hz. 36

3.2 Plots showing (a) mean and (b) variance of MFCC ($C_1 - C_{12}$) of central steady-state portions of vowel /IY/ from signals of different pitch groups: original 100-125 Hz, original 200-250 Hz and pitch transformed versions of original 200-250 Hz pitch group signals with average pitch values transformed to 140-175 Hz pitch range. 37

3.4 State-wise self-loop transition probabilities of the digit 'OH' models corresponding to the adults' data set ADtr and the children's data set CHtr. 41

3.6 Plots of the original linear DFT spectrum along with the smoothed spectra corresponding to MFCC features of a digit 'OH' signal with original and transformed values of (a) OQ (b) RQ (c) SQ. 43

3.7 Distribution of average OQ of the original signals of (a) adults' training set ADtr (b) children's test set CHts1. 44

3.8 Distribution of average RQ of the original signals of (a) adults' training set ADtr (b) children's test set CHts1. 44

3.9 Distribution of average SQ of the original signals of (a) adults' training set ADtr (b) children's test set CHts1. 45

3.10 Plots showing original and frequency warped smoothed Mel spectra for vowel /IY/ having pitch value of around 100 Hz. 45

3.11 Spectrogram of a voiced portion corresponding to word 'Three' extracted from a speech utterance before and after explicit transformation of its average pitch value from 200 Hz to 130 Hz by PSTS method (a) Original 200 Hz (b) Pitch transformed to 130 Hz. The red, green, blue and yellow line plots are the contours of the first, second, third and fourth formants, respectively. 48

3.12 Spectrogram of a voiced portion corresponding to word 'Four' extracted from a speech utterance before and after explicit transformation of its average pitch value from 200 Hz to 130 Hz by PSTS method (a) Original 200 Hz (b) Pitch transformed to 130 Hz. The red, green, blue and yellow line plots are the contours of the first, second, third and fourth formants, respectively. 48

3.13 Distributions of the average pitch of the signals of the children's test set CHts1 after ML-based explicit pitch normalization for all the three pitch groups (as defined in Figure 3.3(b) based on their original pitch values) plotted separately. 50

3.14 Distribution of average speaking rate of signals of the children's test set CHts1 after explicit speaking rate normalization. 52

3.15 Log likelihood distribution of few utterances from the children's test set CHts1 before and after explicit speaking rate normalization on models trained with adults' training set ADtr. 52

3.16 Age group-wise distribution of optimal OQ transformation factors chosen for the signals of the children's test set CHts1. 54

4.1 Plots of the 128-point linear DFT spectra (left panel) and their corresponding cepstra (right panel) for vowel /IY/ having pitch values of around (a) 100 Hz (typical value for male adults' speech) (b) 300 Hz (typical value for children's speech). The peaks in the cepstra corresponding to the pitch harmonics are marked with arrows. Note that for clarity the plots are shown excluding the C_0 coefficient. 64

4.2 Plots of the 30-point uniform filterbank based spectra (left panel) and their corresponding cepstra (right panel) for vowel /IY/ having pitch values of around (a) 100 Hz (b) 300 Hz. The arrow in the cepstrum of the 300 Hz pitch signal shows the location of the first pitch harmonic. 65

4.3 Plots of the 30-point Mel spectra (left panel) and their corresponding cepstra (right panel) for central steady-state portions of vowel /IY/ having pitch values of around (a) 100 Hz (b) 300 Hz. 66

4.4 Plots of the 21-point Mel spectra (left panel) and their corresponding cepstra (right panel) for vowel /IY/ having pitch values of around (a) 100 Hz (b) 300 Hz. 67

4.5 Plots of the signals and the 128-point smoothed Mel spectra (referred to as 'Smoothed') along with their corresponding linear DFT spectra for vowel /IY/ having pitch values of around (a) 100 Hz (b) 220 Hz (c) 300 Hz. 68

4.6 Plots showing 21-point Mel spectra (right panel) of the synthetically generated pitch harmonic spectra (middle panel) corresponding to different pitch frequencies (a) 100 Hz (b) 200 Hz (c) 300 Hz. The synthetic pitch harmonic spectra are created by taking linear DFT of impulse trains shown in corresponding left panel. Note that the slope in the Mel spectra is on account of the outputs of the Mel filters not being normalized by their corresponding areas. 71

4.7 Plots for vowels /AE/ and /IY/ having pitch values of around 100 Hz and 300 Hz (a) Smoothed Mel spectra (b) 13-dimensional truncated MFCCs excluding C_0 (c) relative change in each MFCC for the 300 Hz pitch signal with respect to those for the 100 Hz pitch signal. 72

4.8 Plots of the 128-point linear DFT spectra (left panel), 21-point Mel spectra (middle panel) and their corresponding MFCCs excluding C_0 (right panel) for the synthetically generated pitch harmonic spectra having pitch frequency of around (a) 100 Hz (b) 200 Hz (c) 300 Hz. 73

4.9 Plots showing the relative change in each MFCC ($C_1 - C_{20}$) for the synthetically generated pitch harmonic spectra of different pitch frequencies with respect to those for the synthetic pitch harmonic spectrum having pitch frequency of around 100 Hz (a) 200 Hz (b) 300 Hz. 74

5.1 Plots showing mean (left panel) and bar-plots showing variance (right panel) of each of the coefficients ($C_1 - C_{12}$) of (a) PLPCC features and (b) MFCC features for signals of different pitch groups: 100-125 Hz, 200-250 Hz and 200-250 Hz transformed to 140-175 Hz for vowel /IY/. 81

5.2 Plots of smoothed spectra corresponding to PLPCC features along with the linear DFT spectra for vowel /IY/ having pitch values of around (a) 100 Hz (b) 220 Hz (c) 300 Hz. 82

5.3 Plots for vowel /IY/ having pitch value of around 300 Hz (a) Linear DFT spectrum (b) Smoothed spectra corresponding to PMVDR features computed using various values of LP order. 88

6.1 Plots for vowel /IY/ having pitch value of around 300 Hz (a) Linear DFT spectrum (b) Smoothed Mel spectra computed using various number of filters in the Mel filterbank. 97

6.2 Structures of the Mel filterbank (a) Default (b) Modified. In the modified filterbank the bandwidth of all filters having center frequency below some particular frequency value (say 1 kHz) are modified to have a constant value whereas those of the other filters remain unchanged. 100

6.3 Plots for vowel /IY/ having pitch value of around 300 Hz (a) Linear DFT spectrum (b) Smoothed Mel spectra computed using various bandwidth values for all filters having center frequency below 1 kHz in Mel filterbank. 100

7.1 Plots for vowel /IY/ having pitch value of around 300 Hz (a) Linear DFT spectrum (b) Smoothed Mel spectra corresponding to the base MFCC features of different dimensions. 121

7.2 Graph showing the proposed relation between the length of base MFCC features and the VTLN warp factor for both matched and mismatched test speech signals. 128

7.3 Flow diagram of the proposed algorithm to determine the appropriate length of base MFCC features for recognizing a test speech signal on adults' speech trained models. Here the 'Lookup Table' refers to the proposed relation between the length of base MFCC features and the VTLN warp factor shown graphically in Figure 7.2. 129

7.4 Flow diagram of the proposed algorithm to determine the appropriate length of base MFCC features for recognizing a test speech signal on children's speech trained models. Here the 'Lookup Table' refers to the proposed relation between the length of base MFCC features and the VTLN warp factor shown graphically in Figure 7.2. 131

7.5 Bar graph showing average Bhattacharyya distance across vowel sounds for models trained on adults' and children's speech for connected digit and continuous speech recognition tasks. 133

List of Tables

2.1 Details of the speech corpora used for the speech recognition experiments. 19

2.2 Age group-wise break up of children's speech corpus used in the connected digit recognition task. 20

2.3 Age group-wise break up of children's speech corpus used in the continuous speech recognition task. 20

3.1 Mean and variance of the squared Mahalanobis distances (MD) of MFCC ($C_1 - C_{12}$) features of the original signals of 100-125 Hz and 200-250 Hz pitch groups and the transformed signals with pitch transformation from 200-250 Hz to 140-175 Hz pitch range from the distribution of MFCC features of 75-100 Hz pitch group signals for different vowels. 38

3.2 Mean values of the first four formant frequencies for two different words extracted from a speech signal of the TIDIGITS database before and after its explicit pitch transformation by the PSTS method. 49

3.3 Performance for children's test set CHts1 (with breakup for different pitch groups based on original average pitch values) with and without explicit pitch normalization. The quantity in parentheses shows the number of utterances in that group. The 95% confidence interval for the performance for CHts1 data set is ±0.39 [for $F_o < 250$ Hz, 250 Hz $\leq F_o < 300$ Hz and $F_o > 300$ Hz pitch groups ±0.39, ±0.79 and ±3.37, respectively]. . 49

3.4 Performance for children's test set CHts1 with and without explicit normalization of different acoustic correlates of speech. The 95% confidence interval for the performance for CHts1 data set is ±0.39. 51

3.5 Performance for children's test set PFts with and without explicit normalization of different acoustic correlates of speech. The 95% confidence interval for the performance for PFts data set is ±1.37. 55

3.6 Performance for children's test set CHts1 with and without explicit pitch and speaking rate normalization, VTLN and model adaptation. The numbers given in parentheses are the relative improvements (in %) obtained with respect to their corresponding baseline. 57

3.7 Performance for children's test set PFts with and without explicit pitch and speaking rate normalization, VTLN and model adaptation. The numbers given in parentheses are the relative improvements (in %) obtained with respect to their corresponding baseline. 57

4.1 The center frequencies and the critical bandwidths for human auditory perception as proposed by Zwicker [2] and the center frequencies along with the corresponding bandwidths of all filters of a 21-channel Mel filterbank as per the HTK implementation for 4 kHz signal bandwidth. 69

5.1 Mean and variance of the squared Mahalanobis distances (MD) of the PLPCC (C_1–C_{12}) and MFCC ($C_1 - C_{12}$) (taken from Section 3.2.1 for ease of comparison) features of the original signals of 100-125 Hz and 200-250 Hz pitch groups and the transformed signals with pitch transformation from 200-250 Hz to 140-175 Hz pitch range from the distribution of PLPCC features of 75-100 Hz pitch group signals for different vowels. . 83

5.2 Performance for children's test set CHts1 (with breakup for different pitch groups) with and without pitch normalization for default PLPCC and MFCC (taken from Section 3.3.1.1 for ease of comparison) features. The 95% confidence interval for the performance for CHts1 data set is ±0.39 [for $F_o < 250$ Hz, $250 \leq F_o < 300$ Hz and $F_o \geq 300$ Hz pitch groups ±0.39, ±0.79 and ±3.37, respectively]. 84

5.3 Performance for children's test set CHts1 (with breakup for different pitch groups) with default PMVDR and MFCC features with and without explicit pitch normalization of children's speech. 87

5.4 Performances for children's test set CHts1 (with breakup for different pitch groups) with PMVDR features corresponding to various values of LP orders. 89

5.5 Performances for children's test set CHts1 (with breakup for different pitch groups) using optimized PMVDR features and default MFCC features with and without explicit pitch normalization of children's test speech. 90

6.1 The center frequencies (CF) and the corresponding bandwidths (BW) of all constant-Q filters of filterbank for different number of filters in the Mel filterbank as per the HTK implementation. Note that in order to increase the bandwidth of filters with lower CFs, the BW of filters with higher CFs is also increased and this may result in over-smoothing in higher frequency regions of speech spectrum. 96

6.2 Performances for children's test sets CHts1 (on connected digit recognition task) and PFts (on continuous speech recognition task) with MFCC features computed using various number of filters in the Mel filterbank along with their pitch group-wise breakup. The 95% confidence interval for the performance for CHts1 data set is ± 0.39 [for $F_o <$ 250 Hz, 250 Hz $\leq F_o <$ 300 Hz and $F_o >$ 300 Hz pitch groups ± 0.39, ± 0.79 and ± 3.37, respectively] and for PFts data set is ± 1.37 [for $F_o <$ 250 Hz, 250 Hz $\leq F_o <$ 300 Hz and $F_o >$ 300 Hz pitch groups ± 1.80, ± 1.83 and ± 1.61, respectively]. 98

6.3 Performances for children's test sets CHts1 (on connected digit recognition task) and PFts (on continuous speech recognition task) for various modified constant bandwidth (BW) values for all filters having center frequencies below 1 kHz in the filterbank along with their pitch group-wise breakup. Note that the best ASR performance for each pitch group corresponds to different choice of the bandwidth of filters considered. . . . 101

6.4 Performance for children's test set CHts1 (on connected digit recognition task) with modified constant bandwidth (BW) values of 250 Hz and 300 Hz for various number of filters in the filterbank along with its pitch group-wise breakup. Note that higher pitch groups require modification of the bandwidth of filters up to higher frequencies. 103

6.5 Performance for children's test set PFts (on continuous speech recognition task) with modified constant bandwidth (BW) values of 250 Hz and 300 Hz for various number of filters in the filterbank along with its pitch group-wise breakup. Note that higher pitch groups require modification of the bandwidth of filters up to higher frequencies. 104

6.6 Performances for children's test sets CHts1 (on connected digit recognition task) and PFts (on continuous speech recognition task) using the default MFCC features and MFCC features computed using the proposed pitch normalization algorithm both with and without VTLN and CMLLR. The relative improvement (in %) for each case over its corresponding baseline is given in the parentheses. 107

7.1 Performances for children's test set using MFCC features consisting of varying base feature dimensions on both connected digit recognition and continuous speech recognition tasks. For recognition, the various truncated base MFCC features are also appended with their corresponding first and second order temporal derivatives. 114

7.2 Performances (in descending order) of each of the coefficients of the 39-D default MFCC features for children's test set CHts1 on models trained on adults' speech data set ADtr. 115

7.3 Performances for children's test set CHts1 on models trained on adults' speech data set ADtr using MFCC feature vectors of different dimensions selecting the top d coefficients from the rank-ordered list given in Table 7.2 based on the contribution of each of the coefficients of 39-D MFCC feature vector to the speech recognition performances for the children's test set CHts1 on models trained on adults' speech data set ADtr, referred to as 'Rank Based Feature Selection'. For ease of comparison, the children's ASR performances for CHts1 test set corresponding to MFCC features truncated to different dimensions obtained in Table 7.1 are also given. 116

7.4 Rank ordering of each of the coefficients of the 39-D MFCC features based on their ASR performances for children's test set CHts1 on models trained on adults' speech data set ADtr within the base, the Δ and the $\Delta\Delta$ feature streams. 117

7.5 Variances of squared Mahalanobis distances of MFCC feature vectors of the 'low' (100-125 Hz) and the 'high' (200-250 Hz) pitch group signals from the distribution of MFCC features of 75-100 Hz pitch group signals with different truncations of MFCC features for different vowels. 120

7.6 Performance for children's test set CHts1 on models trained on adults' speech data set ADtr 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 parentheses gives the number of utterances corresponding to that VTLN warp factor. 124

7.7 Performance for children's test set PFts on models trained on adults' speech data set CAMtr for various truncations of base MFCC features along with their VTLN warp factor-wise breakup. 125

7.8 Performance for adults' test set ADts on models trained on adults' speech data set ADtr for various truncations of base MFCC features along with their VTLN warp factor-wise breakup. 126

7.9 Performance for adults' test set CAMts on models trained on adults' speech data set CAMtr for various truncations of base MFCC features along with their VTLN warp factor-wise breakup. 127

7.10 Performances for children's and adults' test sets on adults' speech trained models using default MFCC features (referred to as 'Baseline') and MFCC features derived using the proposed algorithm (referred to as 'Proposed') on both connected digit recognition and continuous speech recognition tasks. 130

7.11 Performances for children's and adults' test sets on children's speech trained models using default MFCC features and MFCC features derived using the proposed algorithm on both connected digit recognition and continuous speech recognition tasks. 132

7.12 Performance for PFts test set using the default MFCC features referred to as 'Default' and MFCC features derived using the proposed algorithm referred to as 'Proposed' both with and without VTLN and CMLLR under both matched and mismatched conditions. 134

List of Acronyms

ASR	Automatic Speech Recognition
BD	Bhattacharyya Distance
BEEP	British English Example Pronunciation
BW	Bandwidth
CF	Center Frequency
CMLLR	Constrained Maximum Likelihood Linear Regression
CMLSN	Constrained MLLR Speaker Normalization
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
EM	Expectation-Maximization
ESPS	Entropic Speech Processing System
HLDA	Heteroscedastic Linear Discriminant Analysis
HMM	Hidden Markov Model
IDCT	Inverse Discrete Cosine Transform
LP	Linear Prediction
LPC	Linear Prediction Coefficients
LPCC	Linear Prediction Cepstral Coefficient
MAP	Maximum *a Posteriori* Adaptation
MD	Mahalanobis Distance
MFCC	Mel Frequency Cepstral Coefficient
ML	Maximum Likelihood
MLLR	Maximum Likelihood Linear Regression
MVDR	Minimum Variance Distortionless Response
OOV	Out of Vocabulary

OQ	Open Quotient
PLP	Perceptual Linear Prediction
PLPCC	Perceptual Linear Prediction Cepstral Coefficient
PMVDR	Perceptual Minimum Variance Distortionless Response
PSOLA	Pitch-Synchronous Overlap-Add
PSTS	Pitch-Synchronous Time-Scaling
RQ	Return Quotient
SAT	Speaker Adaptive Training
SMAPLR	Structural MAP Linear Regression
SQ	Speed Quotient
TD-PSOLA	Time-Domain Pitch-Synchronous Overlap-Add
VTL	Vocal Tract Length
VTLN	Vocal Tract Length Normalization
WER	Word Error Rate

1

Introduction

Contents

1.1 Overview of Automatic Speech Recognition 2
1.2 Challenges in Children's Speech Recognition 3
1.3 Performances of ASR Systems for Children's Speech 6
1.4 Review of Approaches used for Children's ASR 7
1.5 Motivation of the Thesis 13
1.6 Objectives of the Thesis 14
1.7 Organization of the Thesis 14

1.1 Overview of Automatic Speech Recognition

Automatic speech recognition (ASR) refers to the process of transcribing a spoken utterance into its corresponding text through the use of a machine (digital computer). A highly reliable ASR is the need of many rapidly growing application areas such as speech interfaces (increasingly on mobile devices) and indexing of audio/video databases for search. In recent years, significant progress has been achieved in the field of ASR. Nowadays, a number of ASR applications are available commercially but majority of them are developed for adults' speech and work effectively under matched conditions. There is still a considerable gap between human and machine speech recognition performance, particularly in adverse/mismatched conditions [3, 4].

A speech signal is rich in information and it not only contains the linguistic information (the message) but also the information about gender, age, social/regional association, health, and emotional state of the speaker. Besides these inter- and intra-speaker variabilities, a speech signal is also affected by other factors like environmental noises and the characteristics of speech acquisition/channel. The significant variabilities in speech signal caused by these factors make the task of ASR more challenging. The performances of the ASR systems degrade under mismatched conditions i.e., when similar variabilities are not present in either the training or the test speech data. This is the reason why the ASR systems trained with adults' speech perform reasonably well for adults' speech but their performances degrade severely when tested for children's speech [5, 6].

In recent times, because of the potential applications of the ASR systems in education, entertainment and also as an aid to help speech and language development in young children, the children's ASR has increasingly come under investigation [7–10]. Younger children who have yet to develop the dexterity to use the conventional input interfaces such as keypad, mouse and pointers etc. would certainly benefit from a computerized learning aid having a speech interface. It can help children to improve their communication capability and to integrate easily into society. Integrated computer-based multi-modal learning applications help children interact with a virtual tutor to learn to read well [11, 12]. Furthermore, speech recognition technology has been applied to commercial products for children, such as toys and games [6, 13–15]. Examples of major research projects which target children's speech include the STAR project [16], an interactive pronunciation tutor and the LISTEN project [17–19], an interactive reading tutor. A variety of ASR technology enabled applications for children with special needs have been addressed [20, 21]. In [22], the experience acquired using Baldi,

an animated conversational agent, in daily classroom activities with profoundly deaf children is presented. PEAKS is an automatic assessment system of the intelligibility of speech which can be accessed via the internet [23]. This tool allows, for example, to assess speech from children with cleft lip and palate. The use of ASR and spoken dialog technology in the context of other applications including those targeting children with cognitive processing difficulties such as in Dyslexia [24, 25] and Autism Spectrum Disorders [26] is also emerging and offers new challenges for interactive speech technology design and development.

However, the success of the ASR technology lies in developing a robust speech recognition system that can work well irrespective of the usual differences in training and testing conditions. This would avoid the need for building different ASR systems to address different sources of variabilities in speech. Motivated by this, some efforts have been made in this thesis to develop algorithms for improving the children's speech recognition performance on the adults' speech trained ASR systems.

1.2 Challenges in Children's Speech Recognition

All ASR systems comprise of three components: an acoustic model that captures the acoustical properties of the speech sound units (phonemes), a pronunciation dictionary that maps the words into its constituent phonemes and a language model that captures the grammar or syntax of the language of the speech signal. Thus, the properties of a speech signal can be majorally categorized into two types: the acoustic properties and the linguistic properties. Both the acoustic and the linguistic properties of a speech signal are governed by the physiology of the speaker. The control of articulators affects the linguistics in speech while the physical dimensions of vocal tracts and vocal folds have direct influence on the acoustics of speech signal.

Children have large differences in both the acoustic and the linguistic correlates of speech from the adults that make it particularly difficult for ASR systems to deal with children's speech [6, 27, 28]. The various acoustic and linguistic differences include differences in the pitch, the formant frequencies, the average phone duration, the speaking rate, the glottal flow parameters, pronunciation and grammar [29–32]. The children are reported to have different values of mean and variance of the acoustic correlates of speech than those of the adults [27, 33, 34]. For example, the area of the F1-F2 formant ellipses is larger for children than for adults for most vowel phonemes [35]. Another issue in dealing with children's speech is the difficulty to model their constantly changing speech characteristics as all

children undergo rapid development and with varying rates. As children grow, their speech production organs also develop and so their anatomy and physiology keep changing quite significantly. As a result, children's speech has higher inter- and intra-speaker acoustic variabilities than those of adults' speech [28, 30].

1.2.1 Acoustic Correlates of Children's Speech

Several researchers have examined what makes children's speech different from the adults' speech. In a key study [35] by Eguchi and Hirsh, and later summarized by Kent in [36], the age-dependent changes in the formant frequencies and the fundamental frequency measurements of children speakers aged three to thirteen were reported. Children have non-linearly increasing formants located at high values [27, 29, 30, 37, 38]. Also, they have high pitch frequency values causing large spacing between the pitch harmonics [27, 29, 30, 37]. These high formant frequencies and pitch frequency 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 of adult males [27]. In comparison to the presence of 3-4 formants for adults' speech, only 2-3 formants are present for children's speech within 0.3-3.2 kHz frequency range [28]. 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. 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 [27, 29, 30, 39]. The average vowel duration of 5-year old children is reported by Lee *et. al.* in [29] to increase by 36% compared to those of 12-year old children. On analyzing the consonant-vowel transitions in case of adults' and children's speech in [40], 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 [27, 30]. A specific result that is especially relevant for speech modeling is the scaling behavior of formant frequencies with respect to age.

The formant frequencies, the pitch and the speaking rate are perceptually relevant sources of acoustic mismatch. On the other hand, the differences in the voice source parameters due to physiological differences between the speakers affect the source spectrum [41]. The glottal flow of a speech signal

is characterized mainly by the three parameters viz., open quotient (OQ), return quotient (RQ) and speed quotient (SQ). The glottal flow parameters, controlling the shape of the glottal pulse, affect the long-term overall shape (spectral tilt) of the speech power spectrum [41, 42]. For instance, the OQ mainly affects the levels of the lower part of the source spectrum so that a large OQ typically means a higher level of the lowest few harmonics. The RQ affects the steepness of the source spectrum and a large RQ corresponds to greater attenuation of the higher frequencies. These glottal flow parameters i.e., the OQ, RQ and SQ have also been observed to be different for speech corresponding to children and adult speakers [33, 34]. Therefore, the differences in the glottal flow parameters are apparent as differences in the breathiness in speech [43–45]. In [28], young children's speech has been reported to have 60% more breathiness than adults' speech.

Contributing to all this is their increased intra-speaker spectral and temporal variabilities [6,30,40]. Spectral variation, between repetitions of the same vowel by similar speakers, was found to increase by 30% for 5-year old children compared to 12-year old children [27]. The mean Euclidean distance between the cepstral coefficients of the first and second half of vowels for 5-year old children is 20% greater than that for those of 12-year old children. Increase in the intra-speaker spectral and temporal variabilities gives rise to greater overlapping of the phonemic classes making the pattern classification problem even more difficult. It has been reported that children of age of 5 years have about 60% of vowel classification accuracy against that of about 90% of the adults [27].

1.2.2 Linguistic Correlates of Children's Speech

Children 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 [32]. As a result, children's speech have many problems like disfluencies, false-starts and extraneous speech [6, 7, 46]. The frequency of occurrence of mispronunciations for the 8-10 years old children was noted in [46] to be almost twice as high as that 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. The correct inflectional 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's 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 [47]. Linguistic variability in children's speech reduces with age. Older children use simpler linguistic constructs and shorter utterances to convey the intended message. Disfluencies decrease with age and children reach adult-skill level at around 12-13 years of age (somewhat earlier for boys than girls) [47]. So, the ability of the children to use language efficiently to convey the message improves with their age.

1.3 Performances of ASR Systems for Children's Speech

In one of the earlier works on children's ASR reported in [48], Wilpon and Jacobsen studied the children's speech recognition performance using a telephone speech database and a connected digit recognizer. The ASR models were trained using equal amount of speech data from all age groups between 8 and 80. The speech recognition results for speakers of different age groups between 8 and 80 were compared. They showed that the word error rate (WER) for recognition of children's speech is more than 100% higher than that for the adults' speech. Similar observation was also reported by Eskenazi in [49]. The difference in the ASR performance for children in comparison to that for the adults was noted to increase with decreasing age. Similar performances for children's speech have been reported in many other ASR studies as well [5, 6, 9, 50]. Children's speech recognition performances on recognizers trained on adults' speech were noted as WERs of 13% (on digit recognition task) and 49% (on 1000 word vocabulary task) by Potamianos *et. al.* in [51] and Blomberg *et. al.* in [52], respectively. In [53], Gerosa and Giuliani conducted a phone recognition experiment using 28 phone labels. They reported WER of 42% for children's speech on adults' speech trained models.

Also, on exploring the children's speech recognition performances on the children's speech trained models, significant degradation is noted in comparison to those obtained for the adults' speech on the adults' speech trained models. A WER of 23% is reported for children's ASR performance on children's speech trained models by Gerosa and Giuliani in [53]. Using recognizers trained on children's speech, WERs of 7% (for a digit task) and 13% (for a 1000 word vocabulary task) were reported for children's speech by Potamianos *et. al.* in [51] and Blomberg *et. al.* in [52], respectively. These figures represent a 185% and 380% increase in WER for children's speech on adults' speech trained models in comparison to those on children's speech trained models on digit recognition task and on 1000 word vocabulary task, respectively. Thus, children's speech recognition performances are far worse than

those for the adults in general, but is even more degraded on the adults' speech trained models i.e., under the mismatched condition.

Apart from measuring the computer recognition accuracy, in [54], Shona D'Arcy *et. al.* evaluated the human recognition performance as well for adults' and children's speech. They compared the human and machine recognition performance on the same adults' and children's speech data. It is shown that human recognition performance for children's speech exhibits similar effects of age as those observed for automatic systems. This indicates that the effects of age on automatic speech recognition accuracy are due to properties of children's speech rather than the artifacts of the technology.

1.4 Review of Approaches used for Children's ASR

Despite many studies confirming the large acoustic and linguistic differences between the adults' and the children's speech, the relative scarcity of large, publicly-available corpora of children's speech induced researchers to study the possibility to employ speech recognizers trained on adults' speech to decode children's speech. Various approaches that have been explored in literature so far to address different sources of acoustic and linguistic mismatch for improving children's ASR on adults' speech trained models are briefly reviewed in this section.

1.4.1 Acoustic Mismatch

Different methods have been explored in literature for addressing various acoustic differences between adults' and children's speech for improving children's ASR performances. Depending upon the domain in which various acoustic sources of mismatch are addressed, various methods reported in literature can be classified into three broad categories: feature domain, signal domain and model domain approaches.

1.4.1.1 Feature Domain Approaches

Earlier works focused on compensating the acoustic variations induced by differences in the vocal tract lengths, which is one of the major source of acoustic variation between the adults' and the children's speech. Vocal tract length normalization (VTLN) is a speaker normalization method in which the inter-speaker acoustic variability due to different vocal tract lengths across speakers is reduced by warping the frequency axis of the speech spectrum of each speaker [55]. In [56], a strong relationship between the optimal warping factor and the age of the speakers was shown when the

warping factor selection is performed with respect to hidden Markov models (HMMs) trained on adults' speech. A number of studies investigating VTLN for children's ASR show that when a speech recognizer trained on adults' speech is applied to decode children's speech, VTLN is able to significantly improve the children's speech recognition performance [5, 30, 51, 57–61].

In [5], Burnett and Fanty, proposed a rapid approach to perform a speaker-dependent warping of the frequency scale by selecting a Bark offset for each speaker. On adults' speech trained models, they showed a 5.4% and 3.5% improvement in children's ASR performance using a single digit and a seven-digit utterance for adaptation, respectively. As the scaling factors for each of the formant frequencies (F1, F2, F3) are different and are phoneme-dependent, the bi-parametric and phoneme-dependent frequency warping functions were investigated as alternatives to linear frequency warping for speaker normalization in [28]. Additional improvements of 3-5% were reported in children's ASR performance by using a bi-parametric against the linear frequency warping function. Das *et. al.* also carried out frequency warping on a recognizer trained on adults' speech for testing the speech of children from 8 to 13 years of age in context of a command and control application [58]. Speaker independent and speaker dependent frequency warping resulted in an average absolute improvement of 54% and 68%, respectively. In [6], Narayanan and Potamianos reported 45% improvement in the WER for children's ASR on adults' speech trained models by VTLN on a digit recognition task. In comparison to the improvement reported in [6], in [53], Gerosa and Giuliani reported a decrease in the average WER for children's speech from 42% to 33% on an adults' speech trained recognizer after applying VTLN on a phone recognition task. In addition to these, a non-linear extension of VTLN was explored in [59] to derive an optimal filter bank directly from the data for extraction of acoustic features from children's speech.

The use of VTLN has been found to improve the children's ASR performance even on the matched models i.e., the children's speech trained models because of reduction of the inter-speaker variability in children's speech [9, 50]. On applying VTLN to children's speech trained models, in [53], Gerosa and Guiliani reported a decrease in the average WER for children's speech from 23% to 21% on a phone recognition task. Also, in [6], Narayanan and Potamianos have shown a 25% improvement in WER for children's speech recognition on matched models on a digit recognition task.

The acoustic front-end of an ASR system for children is often based on standard Mel frequency cepstral coefficient (MFCC) features. However, few studies show attempts to find out better acoustic

features for children's speech to improve their ASR performance. In [48], the effectiveness of linear prediction cepstral coefficient (LPCC) and MFCC features was compared for children's speech recognition on adults' speech trained models on a connected digit recognition task with telephone speech. Though it was noted that children's ASR performance improves using LPCC features with lower model order, but even greater improvement in the performance was observed using MFCC features. In [62], a special variation in the Mel filterbank, consisting of the normalization of the spectral envelopes using a technique called weighted overlapped spectral averaging was investigated. Using this front-end with adults' and children's speech it was shown that it is more appropriate to assume that the spectral envelopes of any two speakers are linearly scaled version of one another rather than assuming that the whole magnitude spectra including pitch harmonics are scaled. Also, in [38], the length of the analysis window and the width of the filters in the Mel filterbank have been modified to different values for extracting features for children's speech. However, limited effect of these parameters was noted on the children's speech recognition performance. In recent literature, the use of perceptual linear prediction cepstral coefficient (PLPCC) and perceptual minimum variance distortionless response (PMVDR) cepstral coefficient referred to as 'PMVDR' features have also been reported for recognizing children's speech on children's speech trained models [11, 63].

1.4.1.2 Signal Domain Approaches

Among the various sources of acoustic mismatch that have been attributed for degradation in children's ASR performance on adults' speech trained models, in few studies, the differences in the pitch of the signals and the rate of speech have been explicitly normalized in signal domain. In [64], a voice transformation technique has been explored which normalizes the speech signal before being fed to the recognizer. It modifies the speech signal by transforming its pitch using the time-domain pitch-synchronous overlap-add (TD-PSOLA) method and obtaining VTLN by linear compression of the spectral envelope of each window. It is reported that this method reduces the word error rates in the order of 30-45% for children's speech recognition on telephone bandwidth adults' speech trained models. In addition to this, in [59], the speaking rate normalization has been explored to achieve a better ASR performance for children's speech on adults' speech trained recognizer. The speaking rate of each speaker was normalized using the pitch-synchronous overlap-add (PSOLA) algorithm and was shown to give 12% relative improvement in children's ASR performance after rate normalization. Thus, significant improvements have been noted in the children's mismatched ASR performance with

explicit normalization of pitch and speaking rate.

The effect of frequency bandwidth reduction on automatic recognition of children's speech was also investigated in many studies [58,65,66]. In particular, in [65], the children's speech was downsampled from the original 20 kHz to 2 kHz while adults' speech from the original 16 kHz to 2 kHz sampling rate. For each sampling rate a hidden Markov model (HMM) set was trained and then used to recognize the test sets. For children's speech, the decrease in the ASR performance was found to be relatively small down to 6 kHz. A significant degradation in ASR performance was observed between 4 kHz and 2 kHz for both children's and adults' speech, but degradation was much greater for children's speech. It was observed that most values of the third formant for children's speech fall outside telephone bandwidth. This could explain well the low children's ASR recognition performances reported for telephone applications in [48]. Similar effects of bandwidth reduction were also noted on the human recognition performance for children's speech in [54].

1.4.1.3 Model Domain Approaches

Despite various feature domain and signal domain approaches that were explored for children's ASR, the children's speech recognition accuracy was still not as high as that for the adult speakers on the adults' speech trained models [6].

For this reason, general acoustic model adaptation techniques such as maximum *a posteriori* (MAP) [67] adaptation and maximum likelihood linear regression (MLLR) [68] adaptation have also been explored to further improve the children's recognition performance on the adults' speech trained models [60,69]. MLLR applies linear transforms (in the MFCC space) to the entire set of HMMs in order to maximize the likelihood of the adaptation data regardless of whether any examples of the model exist in the adaptation set. On the other hand, MAP uses the generic HMMs as prior knowledge of the parameters of a HMM and combines them with the weighted adaptation data. The weighting depends on the amount of adaptation data available (if no adaptation data exists the generic HMM remains unchanged). In [69], a 15.2% relative improvement has been shown in the ASR performance using MLLR for Italian children's speech on matched models. On the other hand, in [60] relative improvements of 39% and 41% are reported in children's ASR performance on adults' speech trained models using MAP and MLLR adaptations, respectively.

Constrained MLLR speaker normalization (CMLSN) and speaker adaptive training (SAT) have also been studied for improving children's ASR on the adults' speech trained models [30,69]. CMLSN

method transforms the acoustic observation vectors by means of speaker-specific affine transformations obtained through constrained MLLR [30]. A proper scaling factor is used for each speaker or utterance for transforming its corresponding features. SAT performs speaker-specific transformations to compensate for the inter-speaker acoustic variations in the training set [70, 71]. It involves MLLR adaptation of the means of output distributions of continuous density HMMs. In [30], it has been shown that on a continuous speech recognition task relative improvements of 23% and 20% are obtained in children's ASR performance on adults' HMMs using CMLSN and SAT, respectively.

Besides these, the model-space transformation through structural MAP linear regression (SMAPLR) [72] approach has also been explored for improving children's ASR. In [73], the children's ASR performance has been reported to relatively improve by 34% on using SMAPLR adaptation on large vocabulary adults' speech trained ASR models. Significant improvements have been reported in children's ASR performance using these model adaptation techniques on children's speech trained models as well [11, 69, 74].

In order to cope with the age-dependent variability, age-specific modeling of recognizers has also been tried in many studies [48, 50, 51, 57, 58]. Specific models are trained for each target age, or age group of children speakers. Training age-specific speech models requires large amount of data from the target age speakers making the method costlier. So, to reduce the amount of data to be collected for robustly training acoustic models, children are often treated as a homogeneous population group. Acoustic models are trained with speech from children of all ages [50, 51, 58]. However, the recognition performance reported for children's speech is usually significantly lower than that reported for adults' speech on the matched models and it improves as the children's age increases [9, 60, 75, 76]. This correlates well with results of experiments of human perception of speech from children aged 6-11 which have shown that the human word recognition error rate increases as the age of the child decreases [77].

Lately, a different approach was proposed in [78, 79] by considering adults and children as a single population of speakers. Age-independent acoustic models were first conventionally trained by exploiting a small amount of children's speech and a more significant amount of adults' speech. Speaker adaptive acoustic modeling techniques were then used for building ASR system with the unbalanced mixture of adults and children's speech data. The ASR performances for both adults and children were found to be as good as those achieved with age-dependent models. On further using a recognition

vocabulary of 64k words and a tri-gram language model, the WER for children's ASR was noted to be only 24% (relative) higher than the WER for adults' ASR.

However, most of the above studies pointed to the lack of children's acoustic data and resources to estimate speech recognition parameters relative to the over abundance of existing resources for adults' speech recognition. Therefore, many children's speech corpora were later collected for building children's ASR models [49,80,81]. Examples of corpora mostly used for acoustic analysis and modeling are the American English CID children corpus [27], the KIDS corpus [49], the CU Kids' Audio Speech Corpus [9] and the PF-STAR corpus available in the following languages: British English, Italian, German and Swedish [81]. The availability of larger amounts of children's speech data allowed the re-investigation of age-dependent and speaker adaptive acoustic modeling, in the context of medium and large vocabulary children's continuous speech recognition tasks. A noticeable application which makes use of large vocabulary speech recognition for children is presented in [82]. The system with adult and child discrimination capabilities, though addresses users of all ages, makes use of different age-dependent acoustic and language models for adults and children.

1.4.2 Linguistic Mismatch

In addition to building acoustic models customized to children as outlined above, pronunciation and language modeling are also the issues. A standard pronunciation dictionary of an ASR system may result unsuitable for children with a poor pronunciation or for younger children. The importance of using customized language models in recognition of children's speech has been pointed out in several works [6, 58, 76, 83]. It was found that use of a children's speech trained language model substantially improves their recognition performance [58, 84, 85]. The number of tied-states of a speech recognizer was reduced to compensate for the data sparsity by Eskenazi and Pelton in [84]. Also, studies have shown that pronunciation variation has a major influence on the recognition performance for children's speech [31]. The use of a user customized pronunciation dictionary was investigated in [65]. For children who are judged to have good speaking skills, performance is found to be similar to that obtained with adult's speech. However error rates for children whose pronunciation is judged to be poor can be at least four times greater. The results show that ASR performance can be improved by using a customized dictionary, but the improvement is modest.

1.5 Motivation of the Thesis

The large degradation in the children's speech recognition performance on the adults' speech trained models is attributed to both the acoustic and the linguistic characteristics of speech. On a connected digit recognition task as well where the linguistic differences between adults' and children's speech are minimal, the children's ASR performance has been found to be largely degraded in comparison to that of the adults on the adults' speech trained models. This indicates that a large degradation in the children's ASR performance is caused due to the various acoustic differences between adults' and children's speech.

In literature, a number of methods have been used to address various sources of acoustic mismatch for improving the children's speech recognition performance on the adults speech trained models. However, the state-of-art speech recognition performance for children's speech is still largely degraded in comparison to the performance for adults' speech on adults' speech trained models [6]. This indicates that either all acoustic sources of mismatch have not been addressed or have been incompletely addressed for children's speech recognition. Also, all the reported studies have not explored all possible acoustic sources of mismatch in a consistent setup. In particular, besides the formant frequencies and the speaking rate no other acoustic source of mismatch has been explored independently. Therefore, the relative contribution of different acoustic sources of mismatch to the degradation in the children's ASR performance on the adults' speech trained models cannot be ascertained. In addition to this, besides for the formant frequencies, there is not enough explanation available in the literature for how and why the differences in the various acoustic correlates of speech between adults' and children's speech affect the commonly used features and models in case of children's ASR. These facts further take attention when it is noted that, to the best of our knowledge, no study in literature has explored or addressed the pitch in isolation that has already been found to be significantly different for adults' and children's speech for children's ASR.

ASR systems for children's speech widely borrow the architectural choices, approaches and algorithms from state-of-art ASR systems developed to recognize adults' speech. For example, speech signal is often parameterized by MFCC features. However, the standard approaches for feature extraction have in general been optimized for adults' speech. Since there are large differences in various acoustic correlates of speech for adults' and children's speech, the default features used for adults' ASR might not be suitable for parameterizing children's speech. This point was also noted by Wilpon

and Jacobsen who explored efficient features for parameterizing children's speech for their recognition. They compared the effectiveness of linear prediction cepstral features and MFCC features for children's speech recognition on adults' speech trained models on a connected digit recognition task with telephone speech in [48].

1.6 Objectives of the Thesis

In order to improve the performance of children's speech recognition, investigations of speech recognition technologies using features and models suitable for children's speech are needed. Therefore, motivated by the facts discussed in previous section, the objectives of this thesis are:

- To determine the relative significance of each of the already known acoustic sources of mismatch for children's ASR on the adults' speech trained models in a consistent setup. This study is done to identify the significant acoustic sources of mismatch apart from those that are well explored in literature.
- To explore the effect of pitch on the most commonly used MFCC features. This would provide us an insight about the suitable modification that can be incorporated in the MFCC feature computation to efficiently address the pitch mismatch for improving children's ASR on adults' speech trained models.
- To explore the pitch-robustness of other salient ASR features in comparison to MFCC features for children's ASR on adults' speech trained models.
- To develop algorithms for addressing the pitch mismatch in feature domain for improving children's ASR performance on adults' speech trained models.

1.7 Organization of the Thesis

The organization of the rest of this thesis is as follows:

Chapter 2 describes in detail all the methods already reported in literature which have been used in this thesis for various experiments and analysis. The techniques as used in this thesis for modifying various acoustic correlates of speech in a signal are discussed. The standard speaker normalization and the models adaptation techniques used in speech technology are reviewed. Also, the details about

the speech corpora and the experimental setup used in this thesis for training and testing the speech recognition systems are given.

Chapter 3 explores various acoustic sources of mismatch between the adults' and the children's speech for children's speech recognition on adults' speech trained models. The impacts of variations in the acoustic sources of mismatch are studied on the most commonly used MFCC features and on the automatic speech recognition models. Also, the relative significance of those acoustic sources of mismatch is then explored for children's speech recognition on adults' speech trained models in a consistent setup. The acoustic sources of mismatch that are addressed in this study are the formant frequencies, the pitch, the speaking rate and the glottal flow parameters (open quotient, return quotient and speed quotient).

In Chapter 4, a study is done to understand the roles of filterbank and cepstral truncation in removing the pitch-related information in the speech spectrum from the uniform filterbank based and the non-uniform Mel filterbank based spectra and their corresponding cepstra. The Mel cepstra for different pitch signals are then explored to study the cause and the nature of the observed effect of pitch on MFCC features.

In Chapter 5, the pitch-robustness of the salient features that have been reported in literature to perform comparable or better than MFCC for adults' speech recognition are explored for children's speech recognition. The features studied in this work are the PLPCC and the PMVDR features. The effect of pitch variations across speech signals is studied on these features in comparison to that on MFCC features.

An algorithm for normalizing the pitch differences across speech signals during MFCC feature extraction is proposed for children's ASR in Chapter 6. The algorithm modifies the Mel filterbank structure during MFCC feature extraction for each test speech signal based on the average pitch of the test signal. The efficacy of the proposed pitch normalization algorithm is also studied in combination with the existing speaker normalization and model adaptation techniques for children's ASR on adults' speech trained models.

In Chapter 7, MFCC feature truncation is explored for pitch mismatch reduction for children's speech recognition on adults' speech trained models. Based on the observation, an automatic algorithm is proposed for pitch mismatch reduction. The algorithm selects the length of the base MFCC features for recognition of each test speech signal to address its pitch mismatch with respect to the speech

recognition models without any prior knowledge about the speaker of the test utterance. The efficacy of the proposed algorithm is also explored in combination with the existing speaker normalization and model adaptation techniques for children's ASR on adults' speech trained models.

Finally, Chapter 8 summarizes the work presented in this thesis, highlights the main contributions of the work and gives some directions for future research. Note that all speech recognition evaluations in this thesis are done on both connected digit recognition and continuous speech recognition tasks.

2

Speech Corpora and Experimental Setups

Contents

2.1	Introduction	18
2.2	Speech Corpora	18
2.3	Speech Recognition Systems	19
2.4	Methods for Transformation of Acoustic Correlates of Speech	22
2.5	Model Adaptation Techniques	28
2.6	Summary	32

2.1 Introduction

In this chapter, the speech corpora used in this thesis for conducting various ASR experiments on both the connected digit recognition task and the continuous speech recognition task are described. The details of the training and testing of the speech recognition systems used for the connected digit recognition task and the continuous speech recognition task are also given.

Various techniques have been proposed in literature for addressing the acoustic mismatch between the adults' and the children's speech for improving the children's ASR performance on the adults' speech trained models. In this thesis, to determine the relative significance of each of the different acoustic sources of mismatch for children's ASR on adults' speech trained models, those acoustic parameters are transformed using the methods reported in literature. Also, the consistency and efficacy of the pitch normalization approaches proposed in this thesis is explored in conjunction with the various existing speaker normalization and model adaptation techniques. In this chapter, thus, all those methods from literature that have been used in this thesis are also briefly reviewed along with their respective parameter settings kept for the experimental work.

The chapter is organized as follows: Section 2.2 presents the details about the speech corpora followed by the details regarding the speech recognition systems used in this thesis in Section 2.3. A brief review of the methods used for transforming various acoustic correlates of speech in signal domain and in feature domain are explained in Section 2.4. Section 2.5 explains the various model adaptation techniques used in this work for validating the efficacy of our proposed techniques. Finally, the chapter is summarized in Section 2.6.

2.2 Speech Corpora

For, connected digit recognition task, TIDIGITS speech corpus [86] is used for both adults' and children's speech data. For continuous speech recognition task, WSJCAM0 Cambridge Read News corpus [87] is used for adults' speech data and PFSTAR British English corpus [81] is used for children's speech data. The average fundamental frequency (referred to as 'pitch' through out in this thesis) of a speech signal is found by computing the average of the pitch estimates for all frames of the signal. The pitch estimates for all speech frames of a signal are obtained using the Entropic speech processing system (ESPS) tool available in the Wavesurfer [88] software package. All speech data is resampled to 8 kHz. It is well established that children's automatic speech recognition gets more difficult when

Table 2.1: Details of the speech corpora used for the speech recognition experiments.

	Speech Corpus								
	TIDIGITS					WSJCAM0		PFSTAR	
Recognition Task	Connected Digit					Continuous Speech		Continuous Speech	
Sampling Freq.	8 kHz					8 kHz		8 kHz	
Language	American English					British English		British English	
Data Set	ADtr	ADts	CHts1	CHtr	CHts2	CAMtr	CAMts	PFtr	PFts
Purpose	Training	Testing	Testing	Training	Testing	Training	Testing	Training	Testing
Speaker Type	Adults	Adults	Children	Children	Children	Adults	Adults	Children	Children
No. of Speakers	197	81	101	64	49	92	20	122	60
No. of Words	35,566	10,813	25,525	14,725	10,800	132,778	5,320	24,208	5,067
Amount of Data (in hrs.)	5.3	1.6	4.4	2.5	1.9	15.5	0.6	4.8	1.1
Language Model	Equi-probable Wordnet					5k word bi-gram		1.5k word bi-gram	

the speech is sampled at 8 kHz rate due to the loss of some important spectral features of children's speech above 4 kHz frequency range [65, 66]. However, we address this difficult problem because of its potential for applicability to telephone based ASR systems. The details about all these speech corpora that are used in this thesis are given in Table 2.1. The age group-wise details of the children's speech corpora used for connected digit recognition task and continuous speech recognition task are given in Table 2.2 and Table 2.3, respectively.

2.3 Speech Recognition Systems

Throughout this thesis, the ASR performances are evaluated on systems developed using HTK toolkit [89] for two different tasks viz., connected digit recognition task and continuous speech recognition task.

The speech analysis is done using a Hamming window of length 25 ms, frame rate of 100 Hz and a pre-emphasis factor of 0.97. The 13-dimensional MFCC [90] base features ($C_0 - C_{12}$) are computed using a 21-channel filterbank using HTK. In HTK, the Mel filterbank is implemented as a uniform filterbank in Mel frequency domain and then mapped to linear frequency. As a result, in the Mel filterbank implementation in HTK, the bandwidths of the filters up to 1 kHz do not turn out to be of strictly constant value unlike in the implementation proposed by Malcolm Slaney [91]. In addition to the base features, their first and second order temporal derivatives, computed over a span of 5 frames, are also appended making the final features 39-dimensional and henceforth, referred to as the 'default' MFCC features. Cepstral mean subtraction is also applied to all features. The details of the MFCC feature computation process are given in Appendix A.

The word error rate (WER) is used to evaluate the speech recognition performance of various techniques throughout the work in this thesis. The word error rate is computed as follows:

$$\%WER = \frac{\text{Sub} + \text{Del} + \text{Ins}}{\text{Total No. of Words}} \times 100 \tag{2.1}$$

where, 'Sub' represents the number of substitutions, 'Del' represents the number of deletions and 'Ins' represents the number insertions made in the hypothesized text transcript with respect to the true transcription.

Table 2.2: Age group-wise break up of children's speech corpus used in the connected digit recognition task.

	Age Group (Yrs.)				
	6-7	8-9	10-11	12-13	14-15
No. of Speakers (Boys/Girls)	8 (5/3)	31 (12/19)	42 (27/15)	17 (5/12)	3 (1/2)
No. of Utterances	615	2386	3231	1309	231

Table 2.3: Age group-wise break up of children's speech corpus used in the continuous speech recognition task.

	Age Group (Yrs.)				
	4-5	6-7	8-9	10-11	12-13
No. of Speakers (Boys/Girls)	1 (1/0)	12 (5/7)	16 (5/11)	28 (18/10)	3 (3/0)
No. of Utterances	2	20	45	58	4

2.3.1 Connected Digit Recognition

For the connected digit recognition task, the recognizers are developed following the setup described in [92]. The 11 digits (0-9 and OH) are modeled as whole word continuous density HMM using 16 emitting states per word. Each state is a mixture of 5 diagonal-covariance Gaussian distributions with simple left-to-right transitions without any skips over the states. A 3-state model with 6 diagonal-covariance components is used for modeling silence. A single state model with 6 diagonal-covariance components (allowing skip) is used for the short-pause model tied to the center state of the silence model. The details of the procedure used for training and testing a continuous density isolated unit HMM are given in Appendix B. An adults' speech trained recognizer is trained using the adults' speech data set 'ADtr' and is tested against the children's speech data set 'CHts1' and the adults' speech data set 'ADts'. For developing a matched children's ASR system, CHts1 data set which comprises all children's speech data available in TIDIGITS corpus is split into two disjoint sets 'CHtr' (used for training) and 'CHts2' (used for matched testing). The adults' speech recognition performance on children's speech trained models is evaluated using the same adults' speech data set 'ADts'. The baseline recognition performance (in WER) for ADts and CHts1 test sets on the adults' speech trained digit recognizer is 0.43% and 11.37%, respectively. The baseline recognition performance (in WER) for ADts and CHts2 test sets on the children's speech trained digit recognizer is 13.28% and 1.01%, respectively.

2.3.2 Continuous Speech Recognition

For the continuous speech recognition task, the recognizer is developed using cross-word tri-phone acoustic models along with decision tree based state tying. Each tri-phone acoustic model consists of 3 emitting states with 8 diagonal-covariance Gaussian components for each state. A 3-state model with 16 diagonal-covariance Gaussian components is used for modeling silence, and a short-pause model (allowing skip) is constructed with all states tied to the silence model. The adults' speech trained recognizer is trained using the adults' speech data set 'CAMtr' resulting in 2499 tied-states after doing state tying. To evaluate the adults' speech and children's speech recognition performance on this adults' speech trained continuous speech recognizer the 'CAMts' data set and the 'PFts' data set is used, respectively. The children's speech trained recognizer is trained using the children's speech data set 'PFtr' while its recognition performance is evaluated against the children's data set 'PFts' and the

adults' speech data set 'CAMts'. The standard WSJ0 5,000 words closed non-verbalized punctuation vocabulary set and the standard MIT-Lincoln Labs 5k Wall Street Journal bi-gram language model are used for recognition of the adults' test set CAMts having no out of vocabulary (OOV) word. For recognition of the children's test set PFts, a 1,500 words non-verbalized punctuation vocabulary set and a 1.5k bi-gram language model are used. The language model for recognizing children's test set PFts is trained using the transcripts of the children's speech data set PFtr such that the PFts test set has perplexity of 1.02% OOV. The pronunciations for all words are obtained from the British English Example Pronunciation (BEEP) dictionary [87, 93]. The baseline recognition performance (in WER) for CAMts and PFts test sets on the adults' speech trained continuous speech recognizer is 9.92% and 56.34%, respectively. The recognition performance for the children's speech data set PFts is far worse than that obtained on the adults' speech data set CAMts due to the large acoustic mismatch between the adults' training and the children's test data and also due to the loss of spectral information in case of narrowband children's speech. On the children's speech trained continuous speech recognizer, the baseline recognition performance (in WER) for CAMts and PFts test sets is 68.36% and 12.41%, respectively. The poor ASR performance for children's test speech on matched children's speech trained acoustic models in comparison to that for adults' test speech on matched adults' speech trained models is attributed to greater intra- and inter-speaker variability among children than in adults [28, 30]. It is to note that the trend observed in the ASR performances obtained for the above data sets and experimental setups are consistent with that already reported in literature.

The above described recognition systems for both connected digit and continuous speech recognition tasks are used throughout this thesis. This thesis focusses on the children's speech recognition performances on the adults' speech trained models i.e., under mismatched condition. So, unless specified otherwise, all children's speech recognition performances in this thesis refer to children's speech recognition performances on the adults' speech trained models i.e., under mismatched condition.

2.4 Methods for Transformation of Acoustic Correlates of Speech

In this thesis, the various acoustic correlates of speech that have been studied are the pitch, the speaking rate, the formant frequencies and the different glottal flow parameters. For transforming each of these acoustic correlates, the methods already reported in literature have been used. Among these, the pitch, the speaking rate and the glottal flow parameters have been transformed explicitly in

the signal domain in this thesis using the pitch-synchronous time-scaling (PSTS) [1] method. On the other hand, the formant frequencies have been transformed in this thesis in the feature domain using the VTLN technique as proposed by [55]. All these transformation methods used in the following works in this thesis are briefly described in the following subsections.

2.4.1 Signal Domain Method: PSTS

For exploring the effect of different acoustic correlates of speech on the ASR performance, in this thesis, the pitch, the signal duration (for modifying the speaking rate) and the glottal flow parameters viz., the open quotient (OQ), the return quotient (RQ) and the speed quotient (SQ) of the test speech signals are explicitly modified. The PSTS [1] method is reported to provide faithful transformations over a wide range of transformation factors for the above said parameters. In this section, the PSTS method which is used for transforming the average pitch, the signal duration and the glottal flow parameters (OQ, RQ, and SQ) of the speech signals in this thesis is described.

The PSTS method involves pitch-synchronous time-scaling of the linear prediction (LP) residual waveform of the speech signal. By time-scaling the short-time signals, the overlapping interval can be changed maintaining the energy balance of the modified signal. Since the LP residual signal approximates the derivative of the excitation signal, the time-scaling operation also helps in preserving various important parameters of the glottal waveform like the OQ, the RQ and the SQ. Additionally, it also overcomes the problem of energy fluctuations at large pitch modification factors which have been observed in case of pitch transformation using the pitch-synchronous overlap-add based approaches [1].

For doing the pitch-synchronous LP analysis, the pitch marks of the speech signals are extracted. A pitch mark is the location of the short-time energy peak of each pitch pulse in a speech signal. The pitch marks in the voiced regions are computed using the glottal closure instants estimation algorithm in ESPS tool. The algorithm first filters the speech signal with a low and a high band filter and then uses an autocorrelation function to find the pitch mark peaks within the minimum and maximum frequencies specified in the algorithm. In the unvoiced regions of the speech signals, the pitch marks are kept 5 ms equi-spaced. A 10^{th} order pitch-synchronous LP analysis of the speech signal is performed using a 20 ms Hanning window centered on each estimated pitch mark. The residual signal is obtained by inverse filtering the speech signal by a time-varying all-zeros filter defined by the linear prediction coefficients (LPC) associated with each pitch mark. The analysis short-time signals, $x_i(n)$, are obtained by shifting the LP residual signal, $x(n)$, to begin in the previous analysis pitch mark,

$p_a(i-1)$, and then multiplying it with a rectangular window, $rec(n)$, of length equal to the analysis pitch period, $P_a(i) = p_a(i) - p_a(i-1)$:

$$x_i(n) = x(n + p_a(i-1))\, rec(n) \tag{2.2}$$

where,

$$rec(n) = \begin{cases} 1, \ 0 \leq n < P_a(i) \\ 0, \ \text{otherwise} \end{cases} \tag{2.3}$$

2.4.1.1 Transformation of Pitch and Signal Duration

The pitch marks and the LP residual signal are computed as described in Section 2.4.1. The modified pitch mark locations are then computed in accordance to the desired pitch and signal duration (for speaking rate) modification. The shift between successive synthesis pitch marks is equal to the desired pitch period $P_s(j) = p_s(j) - p_s(j-1)$. The synthesis pitch marks $p_s(j)$ are mapped on the estimated analysis pitch marks $p_a(i)$ to determine the short-time signals $x_i(n)$ corresponding to each of the synthesis pitch marks $p_s(j)$.

The pitch and duration transformations can lead to either removal or replication of the analysis short-time signals according to the modification factor. To avoid various phase and frequency discontinuities in the energy envelope and achieve smooth spectral transitions, the non-adjacent l^{th} and r^{th} short-time analysis signals chosen corresponding to two consecutive synthesis pitch marks are first time-scaled to the desired synthesis pitch period $P_s(j)$ to get $z_l(n)$ and $z_r(n)$. The scaling operation is performed with four times over sampling in order to minimize aliasing due to non-ideal low-pass filtering. Then, $z_l(n)$ and $z_r(n)$ are weighted and added to obtain the resultant modified j^{th} short-time synthesis signal $y_j(n)$:

$$y_j(n) = h(n)z_l(n) + h(P_s(j) - n)z_r(n) \tag{2.4}$$

where, $h(n)$ represents a decaying weighting window of size equal to $P_s(j)$, such that $h(n) + h(P_s(j) - n) = 1$. The right half of the Hanning window satisfies this condition.

For pitch transformation factors $N_t > 1$, the speech spectrum gets compressed, giving rise to an "energy hole" at the higher frequencies. Since in our experiments, children's speech is transformed to adults' speech, this problem gets even more enhanced. To overcome this problem, a high frequency

regeneration method is used which regenerates the high-frequency region of the excitation signal, using the correlation between the shape of the glottal flow waveform and the spectrum of the voice source. By pitch-synchronous time-scaling of the open phase of the glottal source waveform, the method reduces the OQ which in turn boosts the energy at the high frequency region of the source spectrum to fill the energy hole. For details of the high frequency regeneration method refer to [94].

Finally, the complete synthesis LP residual signal is obtained by the pitch-synchronous sum of all the synthesis short-time signals using Eqn. (2.5). The modified speech signal is synthesized by passing the modified LP residual through a time-varying all-zeros filter defined by the LPC mapped to the synthesis pitch marks.

$$y(n) = \sum_{j}(y_j(n - p_s(j) + P_s(j))) \tag{2.5}$$

2.4.1.2 Transformation of Glottal Flow Parameters

The original speech waveform is high-pass filtered with a pre-emphasis filter ($\alpha = 0.97$) to attenuate low-frequency rumble and obtain a flatter residual. The pitch marks and the LP residual signal are then computed as described in Section 2.4.1. Corresponding to each pitch cycle a short-time analysis frame is determined using Eqn. (2.3). The following time instants are then estimated for each of the voiced short-time analysis frames:

- Glottal closure instant (n_{cl}): It is estimated as the instant of the first peak after the first zero crossing in the short-time signal.
- Glottal opening instant (n_{op}): It is obtained using the threshold based method described in [95].
- Instant of maximum of the glottal flow (n_p): To compute this instant, first, the DC value between the glottal opening (n_{op}) and every zero crossing to the end of the short-time signal is computed. Then, the zero crossing correspondent to the maximum DC value is chosen as the instant of maximum of the glottal flow.

In order to transform the glottal flow parameters, time-scale transformations are done over the segments corresponding to the glottal flow phases in each of the short-time analysis frames. To avoid aliasing, time-scaling is performed using over sampling with a factor of four. The segments

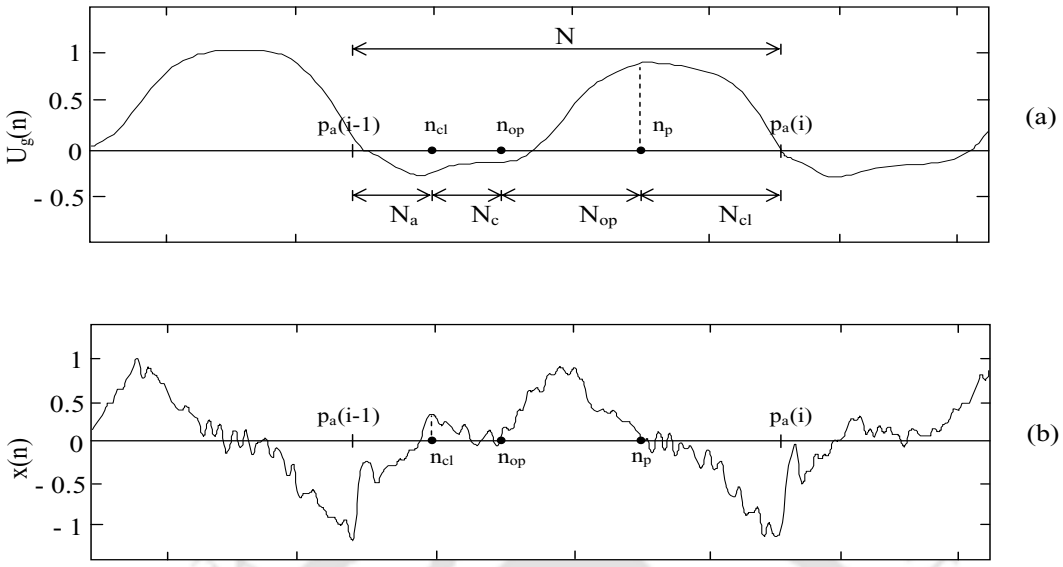


Figure 2.1: Representation of the extracted time instants and the glottal cycle phases in (a) the glottal flow waveform and (b) its time-derivative (i.e., the LP residual signal). The figure is adapted from [1].

corresponding to each of the glottal cycle phases are computed from the extracted time instants using the following relations:

$$\text{Return Phase: } N_a = n_{cl} \tag{2.6}$$

$$\text{Peak Flow Duration: } N_e = N - n_{op} \tag{2.7}$$

$$\text{Closed Phase: } N_c = N - N_a - N_e \tag{2.8}$$

$$\text{Opening Phase: } N_{op} = n_p - n_{op} \tag{2.9}$$

$$\text{Closing Phase: } N_{cl} = N - n_p \tag{2.10}$$

A typical illustration of the various extracted time instants and the glottal cycle phases is given in Figure 2.1.

The open quotient is related to the duration of the open phase and can be expressed as:

$$OQ = (N_a + N_e)/N \tag{2.11}$$

To increase OQ, both the return phase duration and the peak flow duration must be increased. To decrease OQ, both the durations must be shortened. Thus, the time-scale factor is equal to the required modification factor for OQ. Due to time-scale transformation it is necessary to adjust the duration of the closed phase to preserve the pitch period of the glottal waveform as described in [1].

The return quotient is related with the duration of the return phase and determines the cut-off frequency of the spectral tilt. It is computed as:

$$RQ = N_a/N \tag{2.12}$$

The return quotient can be increased or decreased by a time-scale expansion or compression of the return phase. To maintain the pitch period and the open quotient, the peak flow duration is also time-scaled by an adequate factor.

The speed quotient is related to the asymmetry coefficient and accounts for variations in the shape of the segment corresponding to the open phase of the glottal flow. It can be expressed as:

$$SQ = N_{op}/N_{cl} \tag{2.13}$$

The speed quotient can be increased with a time-scale expansion of the opening phase and a time-scale compression of the closing phase so that the peak flow duration $N_e = N_{op} + N_{cl}$ remains constant. SQ can be decreased by the opposite transformation.

Finally, the complete synthesis LP residual signal and the modified synthesis speech signal are computed as described in the Section 2.4.1.1. The sample speech files with the average pitch, the average utterance duration and the average values of the glottal flow parameters (OQ, RQ and SQ) modified by different factors are available at "http://www.iitg.ac.in/ece/emstlab/psts.htm" for assessing the quality of the various transformations.

2.4.2 Feature Domain Method: VTLN

For addressing the mismatch between the adults' and the children's speech due to differences in the formant frequencies, VTLN is used. 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 [96, 97]. In this work, VTLN is performed on an utterance-by-utterance basis on the test speech data as described in [55].

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 [89], 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 [55]:

$$g_\alpha(f) = \begin{cases} \frac{1}{\alpha} f & 0 \le f \le f_0 \\ \frac{1}{\alpha} f_0 + \frac{f_{max} - \frac{1}{\alpha} f_0}{f_{max} - f_0}(f - f_0) & f_0 < f \le f_{max} \end{cases} \tag{2.14}$$

where, f_{max} denotes the maximum signal bandwidth (4 kHz in this work) and f_0 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:

$$\hat{\alpha} = \arg\max_{\alpha} P(X_i^\alpha | \lambda, W_i) \tag{2.15}$$

where, X_i^α represents the warped feature for the i^{th} 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 i^{th} 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 M, thus reducing the inter-speaker acoustic variability between the training and the test data.

2.5 Model Adaptation Techniques

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:

$$Q(M, \hat{M}) = -\frac{1}{2}\sum_{r=1}^{R}\sum_{m_r=1}^{M_r}\sum_{t=1}^{T} L_{m_r}(t)[K^{(m_r)} + log(|\hat{\mathbf{\Sigma}_{m_r}}|) + (\mathbf{o}(t) - \hat{\mu_{m_r}})^T \hat{\mathbf{\Sigma}_{m_r}^{-1}}(\mathbf{o}(t) - \hat{\mu_{m_r}})] \quad (2.16)$$

where, M represents the current recognition model set, $\hat{M}$ represents the adapted model set, r denotes the training observation sequence from a set of training data observations where $1 \leq r \leq R$. m_r denotes the m^{th} mixture component of the r^{th} observation sequence while M_r denotes the total number of mixture components of the r^{th} observation sequence. $\mathbf{O}$ represents the sequence of d-dimensional observations, $\mathbf{o}(t)$ denotes the observation at time t where $1 \leq t \leq T$. μ_{m_r} represents the mean vector for the mixture component m_r, $\mathbf{\Sigma}_{m_r}$ represents the covariance matrix for the mixture component m_r and $K^{(m_r)}$ subsumes all constants. $L_{m_r}(t)$ represents the occupancy probability for the mixture component m_r at time t and is defined as,

$$L_{m_r}(t) = P(q_{m_r}(t)|M, \mathbf{O}_T) \quad (2.17)$$

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.

In this work, all model adaptations are performed as supported in the HTK computing speaker-specific transformations in all cases.

2.5.1 MLLR

In MLLR [68] model adaptation technique, a set of linear transformations for the mean μ and variance $\mathbf{\Sigma}$ parameters of the Gaussian distributions $\mathcal{N}(\mathbf{x}; \mu, \mathbf{\Sigma})$ 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. In this work, the effect of adapting the mean and the variance parameters of the models is studied separately.

2.5.1.1 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 in this thesis as 'MLLR-MEAN'. The adapted $d \times 1$ mean vector $\hat{\mu}$ is given by,

$$\hat{\mu} = W\xi \quad (2.18)$$

where, W represents the $d \times (d+1)$ transformation matrix (where, d is the dimensionality of the data) and ξ represents the extended mean vector.

$$\xi = [w\ \mu_1\ \mu_2\ ...\ \mu_d]^T \tag{2.19}$$

where, w represents a bias offset which is kept as 1 (default value within HTK) in this work and μ represents a $d \times 1$ mean vector. Hence, W can be decomposed into

$$W = [\mathbf{b}\ A] \tag{2.20}$$

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

2.5.1.2 MLLR-COV

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

$$\hat{\mathbf{\Sigma}} = H\mathbf{\Sigma}H^T \tag{2.21}$$

where, 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:

$$\mathcal{N}(\mathbf{o}; \mu, H\mathbf{\Sigma}H^T) = \frac{1}{|H|}\mathcal{N}(H^{-1}\mathbf{o}; H^{-1}\mu, \mathbf{\Sigma}) = |A|\mathcal{N}(A\mathbf{o}; A\mu, \mathbf{\Sigma}) \tag{2.22}$$

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

2.5.2 CMLLR

In constrained MLLR (CMLLR) feature adaptation technique, a set of linear transformations for the features are estimated so as to modify the feature vectors such that their likelihoods increase with respect to the given model.

In [71], it is shown that mean μ and variance $\mathbf{\Sigma}$ of a Gaussian density $\mathcal{N}(\mathbf{o}; \mu, \mathbf{\Sigma})$ associated with a HMM state can be adapted by means of an affine transformation, estimated in the maximum likelihood framework, in the following way:

$$\hat{\mu} = \tilde{A}\mu + \tilde{\mathbf{b}}, \quad \hat{\mathbf{\Sigma}} = \tilde{A}\mathbf{\Sigma}\tilde{A}^{*} \tag{2.23}$$

where, $\tilde{A}$ and $\tilde{\mathbf{b}}$ represent the matrix and the offset vector of the so-called constrained model-space transformation [71]. 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 [71] using the following identity:

$$\mathcal{N}(A\mathbf{o} + \mathbf{b}; \mu, \mathbf{\Sigma}) = |A^{-1}|\mathcal{N}(\mathbf{o}; A^{-1}(\mu - \mathbf{b}), A^{-1}\mathbf{\Sigma}A^{-1^{*}}) \tag{2.24}$$

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

$$\tilde{A} = A^{-1}, \quad \tilde{\mathbf{b}} = -A^{-1}\mathbf{b} \tag{2.25}$$

Thus, the adapted $d \times 1$ observation vector $\hat{\mathbf{o}}$ is given by,

$$\hat{\mathbf{o}} = W_o\ \zeta \tag{2.26}$$

where, W_o represents the $d \times (d + 1)$ transformation matrix (where, d is the dimensionality of the data) and ζ represents the extended observation vector.

$$\zeta = [w\ o_1\ o_2\ ...\ o_d]^T \tag{2.27}$$

where, w represents a bias offset which is kept as 1 (default value within HTK) in this work and $\mathbf{o}$ represents a $d \times 1$ observation vector.

Hence, W_o can be decomposed into

$$W_o = [\mathbf{b}\ A] \tag{2.28}$$

where, A represents a $d \times d$ 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 in the likelihood calculation.

$$\mathcal{L}(\mathbf{o}; \mu, \mathbf{\Sigma}, A, \mathbf{b}) = |A|\mathcal{N}(A\mathbf{o} + \mathbf{b}; \mu, \mathbf{\Sigma}) \tag{2.29}$$

2.6 Summary

In this chapter, the details of the speech corpora and the experimental setup used for various ASR performance evaluations in this thesis are described. The speech corpora that are used further in this thesis for adults' and children's speech recognition experiments are TIDIGITS, WSJCAM0 and PFSTAR. The TIDIGITS speech corpus is used to perform ASR experiments on connected digit recognition (limited vocabulary) task while WSJCAM0 and PFSTAR speech corpora are used for conducting ASR experiments on continuous speech recognition task. In addition to this, the VTLN approach as used in this thesis for normalization of the acoustic mismatch due to differences in the formant frequencies is described. The PSTS method, reported in literature, is then reviewed along with its experimental settings as used in this thesis for transformation of pitch, speaking rate and glottal flow parameters. The model adaptation techniques already existing in literature viz., MLLR and CMLLR are also discussed in brief which are used in the upcoming chapters for exploring the efficacy of the methods proposed in this thesis.

3

Role of Various Acoustic Sources of Mismatch in Children's ASR

Contents

3.1	Introduction	34
3.2	Effect of Various Acoustic Sources of Mismatch on MFCC Features & ASR Models	35
3.3	Relative Significance of Various Acoustic Sources of Mismatch for Children's ASR	46
3.4	Combining VTLN and Explicit Acoustic Normalization with Model Adaptation	56
3.5	Summary	58

3.1 Introduction

In literature, various acoustic sources of mismatch between adults' and children's speech have been attributed to the degradation in the children's ASR performance on the adults' speech trained models [6, 27, 30, 31]. A number of studies have explored different feature domain and model domain techniques for addressing various acoustic sources of mismatch for improving children's ASR performance [30, 59, 64]. However, besides for the formant frequencies, there is still not enough explanation available in literature for how the differences in the various acoustic correlates of speech affect the commonly used features and models in case of children's ASR. Also, all the reported studies in literature have not explored all possible acoustic sources of mismatch in a consistent setup. As a result, the relative contribution of different acoustic sources of mismatch to the degradation in the children's ASR performance on the adults' speech trained models cannot be still ascertained.

Motivated by these, in this chapter, the effect of differences in various acoustic sources of mismatch between the adults' and the children's speech, which have already been identified in literature, is explored on the most commonly used MFCC [90] features and HMM-based ASR models in context of children's ASR on adults' speech trained models. The acoustic correlates that are included in this study are the pitch, the speaking rate, the formant frequencies and the glottal flow parameters: open quotient (OQ), return quotient (RQ) and speed quotient (SQ). Following this, an effort is made to determine the relative significance of each of these acoustic correlates for the children's ASR on the adults' speech trained models by explicitly normalizing them using the techniques available in literature. The study is done initially on a limited vocabulary connected digit recognition task where the linguistic differences between the adults' and the children's speech would be minimal. Further, to verify the generality of the observations made on a digit recognition task, the study is also repeated on a continuous speech recognition task. To study the effect of only the acoustic mismatch between the adults' and the children's speech, the respective language models are used for the adults' and the children's speech recognition on a continuous speech recognition task.

The rest of the chapter is organized as follows. In Section 3.2, the effect of differences in various acoustic correlates of speech is explored on the commonly used MFCC features and HMM-based ASR models. The children's ASR performances with explicit normalization of various acoustic correlates are given in Section 3.3. In Section 3.4, combination of the explicit acoustic normalization with the existing model adaptation techniques is explored. Finally, this chapter is summarized in Section 3.5.

3.2 Effect of Various Acoustic Sources of Mismatch on MFCC Features & ASR Models

In this section, the effect of differences in various acoustic sources of mismatch between the adults' and the children's speech is explored on MFCC features and HMM-based acoustic models. The various acoustic correlates that are studied are the pitch, the speaking rate, the formant frequencies and the glottal flow parameters (OQ, RQ and SQ). To study the effect of various acoustic sources of mismatch on MFCC features, the vowel speech data derived from the TIMIT [98] database is used. The extracted vowel speech data is downsampled from 16 kHz to 8 kHz in order to be consistent with the recognition experiments reported throughout in this thesis.

3.2.1 Pitch

The Mel filterbank is employed during MFCC feature computation so as to smooth out the pitch harmonics in the speech spectrum. Thus, MFCC features are expected to predominantly capture the envelope of the speech spectrum and be devoid of any pitch-related information. However, in [99], it has been reported that the pitch can be predicted from MFCC features and thus, has shown the possibility of reconstructing a speech signal from its MFCC features. Also, improvement in the phone classification performance has been observed with pitch-dependent normalization of Mel cepstrum [100]. In context of children's ASR using MFCC features, a study has reported an improved children's ASR performance on reduction of the pitch of the signals [64]. Recently, in [101], the pitch adaptive MFCC features have been shown to improve the adults' ASR performance on matched models on large vocabulary ASR tasks. Motivated by these studies, we explore the effect of pitch on MFCC features.

First, the effect of pitch is explored on the smoothed Mel spectrum corresponding to the 13-dimensional (13-D) base MFCC ($C_0 - C_{12}$) features. The smoothed Mel spectrum corresponding to MFCC is derived by computing a 128-point inverse discrete cosine transform of the 13-D base MFCC features after appending 115 zeros to the 13-D MFCC features. Figure 3.1 shows the smoothed Mel spectra along with the linear discrete Fourier transform (DFT) spectrum for central steady-state portions of vowel /IY/ having pitch values of around 100 Hz, 220 Hz and 300 Hz. Some significant distortions are observed in the smoothed Mel spectral envelope particularly at the lower frequencies (below 1 kHz) for 220 Hz and 300 Hz pitch signals when compared with that of the 100 Hz signal.

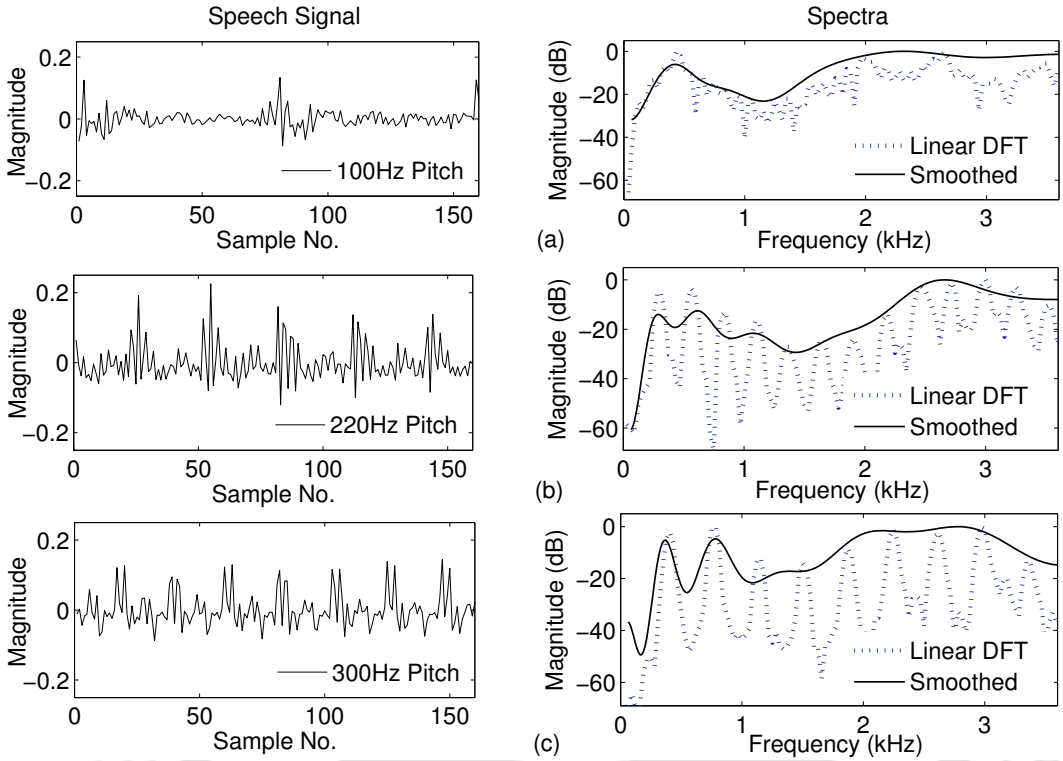


Figure 3.1: Plots of the signals and the smoothed Mel spectra (referred to as 'Smoothed') along with their corresponding linear DFT spectra for central steady-state portions of vowel /IY/ having pitch values of around (a) 100 Hz (b) 220 Hz (c) 300 Hz.

Motivated by these, we further analyze the effect of the pitch variations among the speech signals on MFCC ($C_0 - C_{12}$) features. The speech signals in the TIMIT database belonging to 'low' (100-125 Hz) and 'high' (200-250 Hz) pitch groups are selected. The average pitch (F_o) of the signals is estimated using the ESPS tool available in the Wavesurfer software package [88] as described in Section 2.2. The central steady-state portions of 7 different vowels present in the selected signals are then extracted and their corresponding MFCC features are computed. For each vowel, approximately 2000 frames corresponding to the central steady-portions are used for the study.

The plots of mean and the bar-plots showing variance of each of the 12 dimensions ($C_1 - C_{12}$) of MFCC features for signals belonging to 'low' and 'high' pitch groups for a representative vowel /IY/ are shown in Figure 3.2(a) and Figure 3.2(b), respectively. It is noted that the variances of the higher order coefficients of MFCC features of the high pitch group signals are much larger than those corresponding to the low pitch group signals. To validate whether these differences in the variances of MFCCs are caused by the differences in the average pitch values of the signals in the two groups, the pitch of the 200-250 Hz pitch group signals is transformed to 140-175 Hz pitch range through a

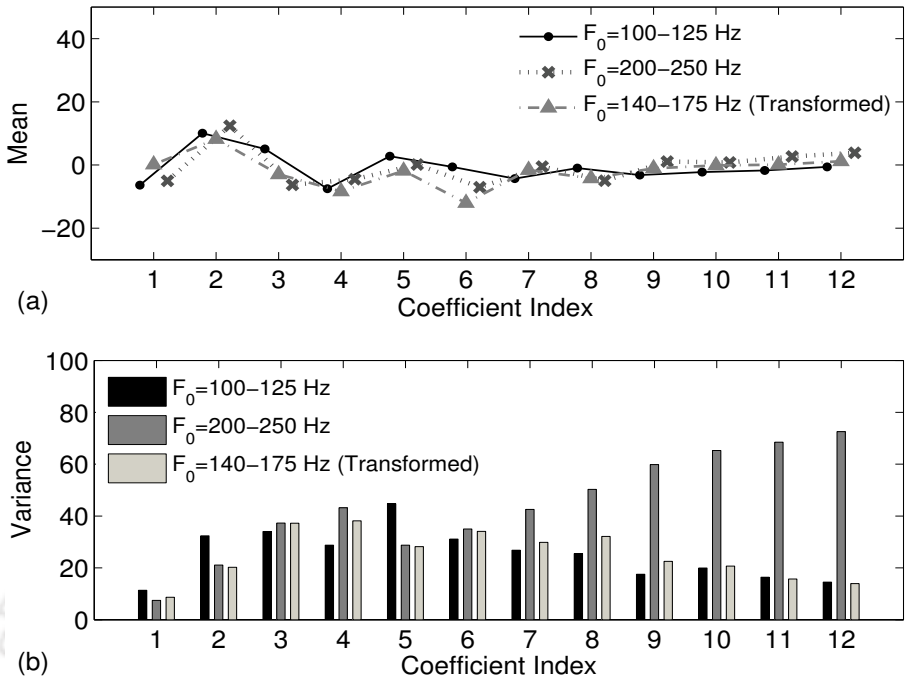


Figure 3.2: Plots showing (a) mean and (b) variance of MFCC ($C_1 - C_{12}$) of central steady-state portions of vowel /IY/ from signals of different pitch groups: original 100-125 Hz, original 200-250 Hz and pitch transformed versions of original 200-250 Hz pitch group signals with average pitch values transformed to 140-175 Hz pitch range.

constant factor of 0.7 for all vowels using the PSTS [1] method as described in Section 2.4.1.1. A little higher transformation factor of 0.7 is used in place of 0.5 so as to limit the possible distortions that might appear during explicit pitch transformation process with the use of lower transformation factors. The plots of mean and the bar-plots showing variance of each of the 12 dimensions ($C_1 - C_{12}$) of MFCC features for signals belonging to the transformed pitch group for vowel /IY/ are also shown in Figure 3.2(a) and Figure 3.2(b), respectively. It is noted that on reduction of pitch of the signals the variances of the higher order coefficients of their MFCC features also reduce considerably. The cause of the occurrence of some distortions in the smoothed Mel spectra of high pitch signals and the increase in the variances of the higher order MFCCs with increase in the pitch of the signals is further explained in detail in Chapter 4.

In continuous density HMMs, the likelihood computation for recognition purpose involves the use of the Mahalanobis distance (MD) measure [102]. Thus, to quantify the effect of these variations in MFCC features due to changes in average pitch across the signals on the classification performance of the speech recognition models, the MD measure is employed. The greater is the MD of a feature vector with respect to a given recognition model, the poorer is the classification and thus, the recognition

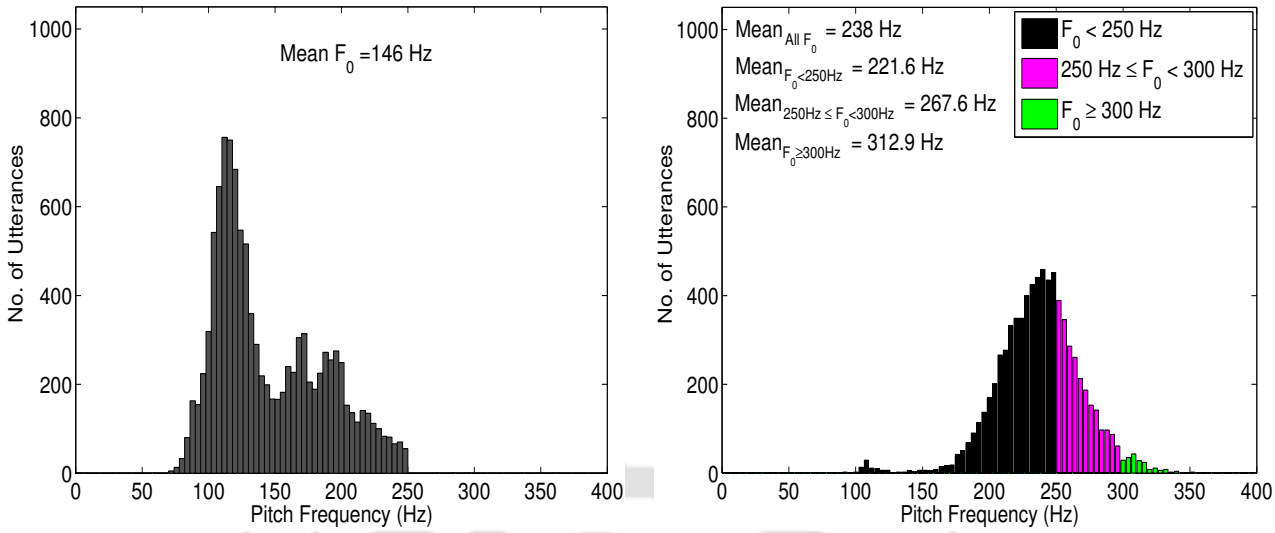
Table 3.1: Mean and variance of the squared Mahalanobis distances (MD) of MFCC ($C_1 - C_{12}$) features of the original signals of 100-125 Hz and 200-250 Hz pitch groups and the transformed signals with pitch transformation from 200-250 Hz to 140-175 Hz pitch range from the distribution of MFCC features of 75-100 Hz pitch group signals for different vowels.

Pitch Group	Squared MD (Mean/Variance)	
	Vowel /AE/	Vowel /IY/
100-125 Hz (Original)	12.4 / 60.2	13.5 / 85.3
200-250 Hz (Original)	65.4 / 2553.1	59.2 / 1735.4
200-250 Hz (Transformed to 140-175 Hz)	35.3 / 282.7	34.7 / 246.8

performance of the model for that feature. The MD is computed for the MFCC ($C_1 - C_{12}$) feature vectors of all signals of the low pitch group, the high pitch group and the pitch transformed version of the high pitch group to 140-175 Hz pitch range with respect to the distribution of MFCC features of the 75-100 Hz pitch group signals using Eqn. 3.1 :

$$\texttt{MD}(\mathbf{x}, \mu_L) = \sqrt{(\mathbf{x} - \mu_L)^T \mathbf{\Sigma}_L^{-1} (\mathbf{x} - \mu_L)} \tag{3.1}$$

where, $\mathbf{x}$ represents the feature vector whose distance is to be computed. μ_L represents the mean and $\mathbf{\Sigma}_L$ represents the diagonal-covariance of the distribution of MFCC features of 75-100 Hz pitch group signals. The mean and variance of the squared MD of the feature vectors of all signals of a pitch group are then computed in order to determine the average distance and the extent of distance of the feature vectors of that pitch group from the 75-100 Hz pitch group distribution, respectively. This procedure is performed on all pitch groups separately. The means and variances of squared MD of MFCC ($C_1 - C_{12}$) features of signals of different pitch groups for vowels /AE/ and /IY/ are given in Table 3.1. It is noted that the mean and variance of squared MD are greater for the high pitch group than those for the low pitch group. Thus, the models trained with the 75-100 Hz pitch group signals would have poor classification for the original high pitch group signals in comparison to the low pitch group signals. However, on reduction of the pitch of the high pitch group signals, the mean and variance of squared MD for the high pitch group signals are significantly reduced in comparison to those corresponding to the original high pitch group signals. This indicates that increase in the pitch of the signals increases the MD of their MFCC features from the given models. This would in



(a) (b)

Figure 3.3: Distribution of average pitch of the original signals of (a) adults' training set ADtr (b) children's test set CHts1. Three broad pitch groups have been marked with three different colors for studying their distribution after explicit pitch normalization.

turn cause degradation in the classification performance and therefore, the recognition performance of the given ASR models for those signals.

The difference in the average pitch values of the children's and the adults' speech data used in this work can be understood by observing the distribution of the average pitch of the signals of the adults' training set ADtr and the children's test set CHts1 as shown in Figure 3.3(a) and Figure 3.3(b), respectively. It is noted that the mean of the pitch distribution of the children's test set CHts1 is nearly of the order of 1.6 to that of the adults' training set ADtr. The bi-modal distribution of the average pitch of the adults' training set ADtr is attributed to the presence of both male and female speakers in that data set. Thus, as expected, children have significantly higher pitch values than adults. Therefore, on account of the already noted effect of pitch of the signals on the MD of their features, children's speech recognition would be significantly degraded on the adults' speech trained models. Also, it is hypothesized that the reduction of the variances of MD with explicit pitch normalization on account of reduction of the variances of higher order coefficients of MFCC feature would in turn result in improvement in the ASR performance for children's speech.

3.2.2 Speaking Rate

In literature, many studies have reported that the variation in the speaking rate affects the acoustic patterns of speech by restructuring the relationship between the acoustic cues and the phonetic categories and thus, affect both perception and production of phones [103–105]. Also, some studies have reported significant degradation in the ASR performance for exceptionally fast and slow speaking rate adults' speech [106–109]. The speaking rate has been observed to affect the duration of vowels the most [110,111]. The ratio of duration between a consonant and a vowel of a CV-syllable in a fast speech is kept almost the same as that in a neutral speech while vowel lengthening becomes significantly large in the slow speech [111]. Following that, in [106], more explicit modeling of the phone durations by modification of the state-transition probabilities of the HMM-based acoustic models has been explored for improving the adults' ASR performance. In [107], it was noted that the increase in the phone exit probability in each state of the model significantly improved the recognition performance for adults' speech having fast speaking rates. Therefore, the models trained with fast and slow speaking rate speech data would have largely different transition probabilities. From literature, it is already known that children have lower speaking rate than adults [27]. Also, in [59], improvement in the children's ASR performance has been noted on adults' speech trained models after normalization of speaking rates of children's speech. Therefore, motivated by these, we explore the effect of different speaking rates on the state-transition probabilities of adults' and children's speech trained HMM-based acoustic models.

In order to explore the effect of differences in adults' and children's speaking rate on the state-transition probabilities of ASR models, the self-loop transition probabilities of each of the 16 emitting states of the single digit 'OH' models corresponding to the children's speech data set CHtr (less speaking rate) and the adults' speech data set ADtr (high speaking rate) shown in Figure 3.4 are compared. The digit 'OH' models corresponding to both the adults' training set ADtr and the children's training set CHtr are generated following the setup described in Section 2.3.1. It is noted that the children's speech trained model has greater self-loop transition probability across all states in comparison to that of the adults' speech trained model. This implies that in case of children's speech each observation has more probability to remain in same state for longer time than in case of adults' speech. This is attributed to the fact that children's speech has lesser speaking rate and thus have longer sentence/phone durations in comparison to those for the adults' speech.

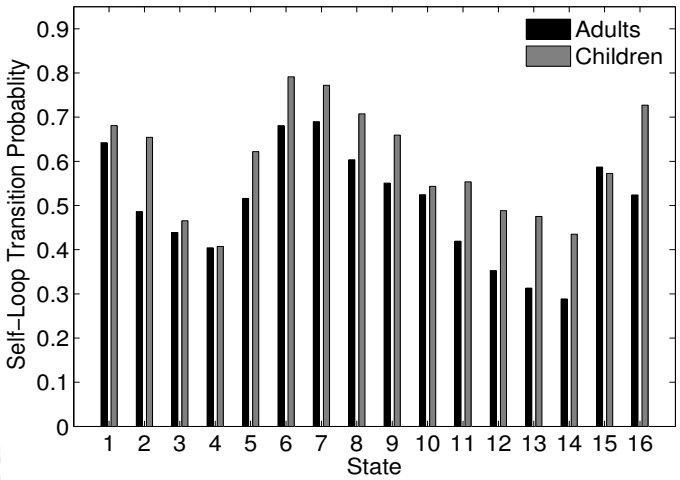
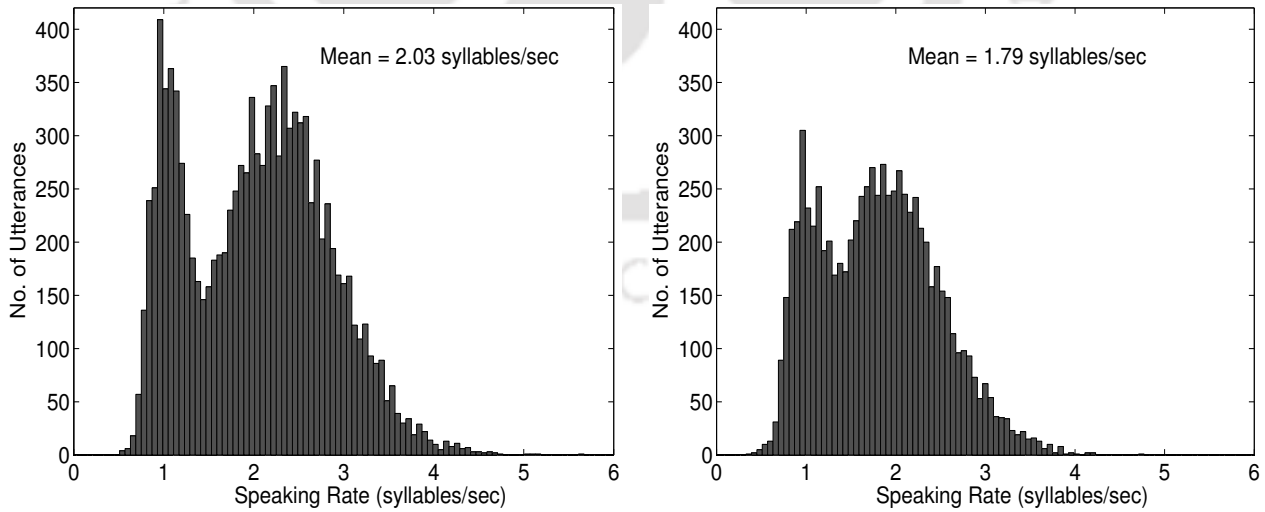


Figure 3.4: State-wise self-loop transition probabilities of the digit 'OH' models corresponding to the adults' data set ADtr and the children's data set CHtr.

The difference in the average speaking rate of the children's and the adults' speech data used in this work can be understood by observing the distribution of the average speaking rate of the adults' training set ADtr and the children's test set CHts1 as shown in Figure 3.5(a) and Figure 3.5(b), respectively. It is noted that the mean of the speaking rate distribution of the adults' training set ADtr is 1.2 times that of the children's test set CHts1. Thus, as expected, children have longer sentence duration, and thus, lower speaking rate than adults.



(a) (b)

Figure 3.5: Distribution of average speaking rate of the original signals of (a) adults' training set ADtr (b) children's test set CHts1.

Therefore, on account of the already noted effect of rate of speech of training data on the state-transition probabilities of the models, the children's ASR performance on the adults' speech trained models would be adversely affected. So, it is hypothesized that explicit speaking rate normalization of children's speech might improve their recognition on the adults' speech trained models which have comparatively fast speaking rate.

3.2.3 Glottal Flow Parameters

The voice source signal and its corresponding source spectrum are found to differ significantly among speakers due to differences in their physiological attributes [41, 42]. For instance, the OQ mainly affects the levels of the lower part of the source spectrum so that a large OQ typically means a higher level of the lowest few harmonics. The RQ affects the steepness of the source spectrum so that a large RQ corresponds to greater attenuation of the higher frequencies. As a result, these glottal flow parameters govern the voice quality and thus, the breathiness in the speech [43–45]. In literature, young children have been reported to have 60% more breathiness in their speech than adults [28]. In addition to this, the glottal flow parameters like OQ, RQ and SQ have also been observed to be different for adults' and children's speech [33,34]. Therefore, motivated by these observations reported in literature, the effect of variations in each of the glottal flow parameters (OQ, RQ and SQ) on MFCC features is studied.

In order to explore the effect of differences in these three chosen glottal flow parameters on MFCC features, the smoothed Mel spectra corresponding to 13-D MFCC ($C_0 - C_{12}$) of a central steady-state frame obtained from a digit 'OH' speech signal before and after transformation of each of these glottal flow parameters by different factors are compared as shown in Figure 3.6. It is noted that changes in the values of any of these three glottal flow parameters give rise to some changes in the smoothed Mel spectrum corresponding to MFCC.

The differences in the average values of the three glottal flow parameters of the children's and the adults' speech data used in this work can be understood by observing the distribution of the average values of the OQ, RQ and SQ for both the adults' training set ADtr and the children's test set CHts1 as shown in Figure 3.7, Figure 3.8 and Figure 3.9, respectively. It is noted that the mean values of all the three glottal flow parameters for the adults' training set ADtr are smaller than those for the children's test set CHts1. Since higher values of these parameters indicate more breathiness, as expected, children have more breathiness in their speech than adults. Some changes have already been observed in the

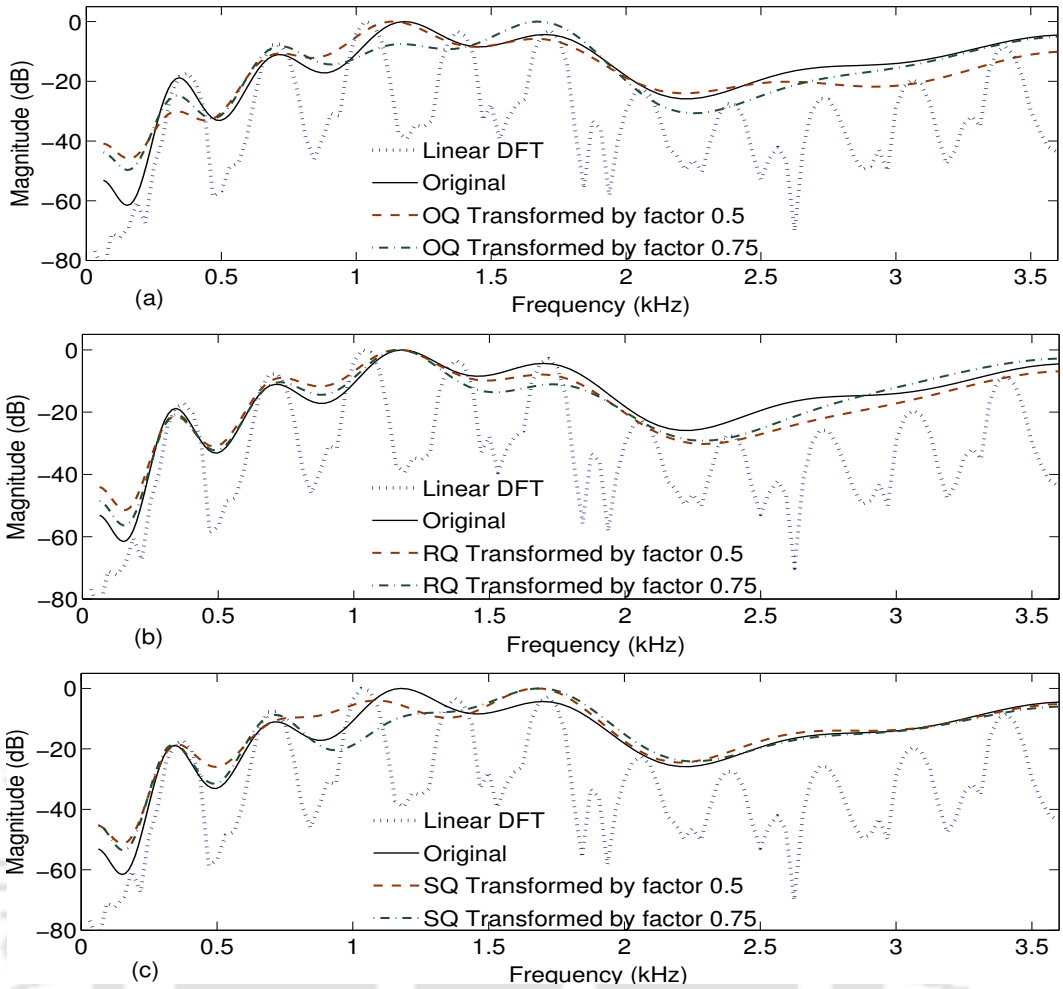


Figure 3.6: Plots of the original linear DFT spectrum along with the smoothed spectra corresponding to MFCC features of a digit 'OH' signal with original and transformed values of (a) OQ (b) RQ (c) SQ.

smoothed Mel spectra corresponding to MFCC features of signals after transformation of their glottal flow parameters to different values. Therefore, the effect of explicit normalization of these glottal flow parameters for the children's speech signals is further explored on their ASR performance.

3.2.4 Formant Frequencies

The differences in the vocal tracts across speakers cause changes in the formant locations and the formant bandwidths in their speech spectrum [112]. In literature, the VTLN has been well explored for reducing the mismatch due to differences in the formant frequencies between the adults' and the children's speech for improving children's ASR on adults' speech trained models [6, 30]. It diminishes the effect of differences in the vocal tract length among different speakers by appropriately warping the frequency axis of the speech spectrum during signal analysis [97]. As a result, the acoustic mismatch

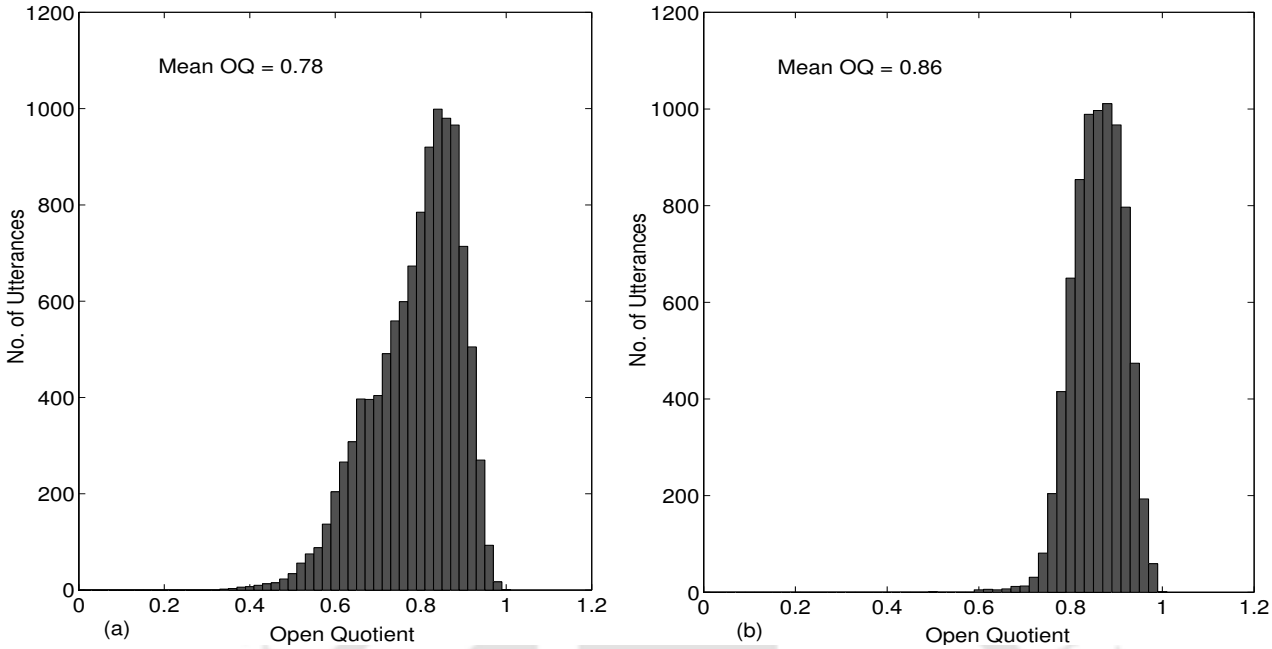


Figure 3.7: Distribution of average OQ of the original signals of (a) adults' training set ADtr (b) children's test set CHts1.

between the test speech signal and the speech recognition models on account of the differences in the formant frequencies of the test signal and the training set signals is reduced. Role of frequency warping

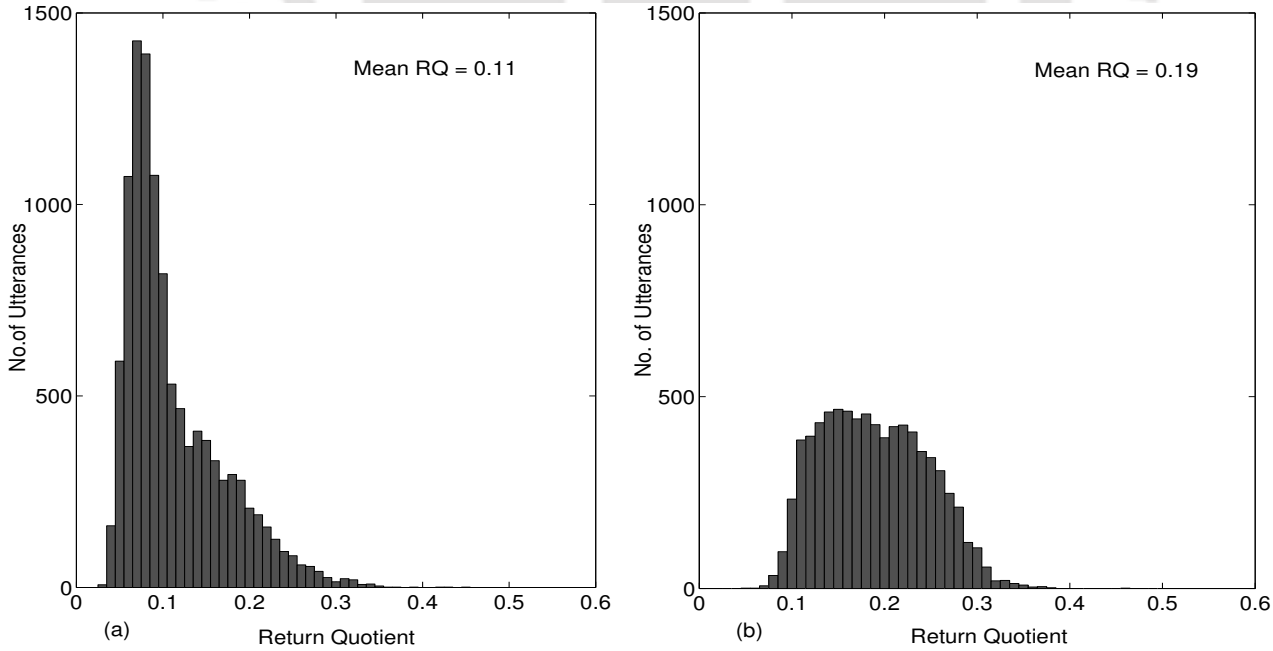


Figure 3.8: Distribution of average RQ of the original signals of (a) adults' training set ADtr (b) children's test set CHts1.

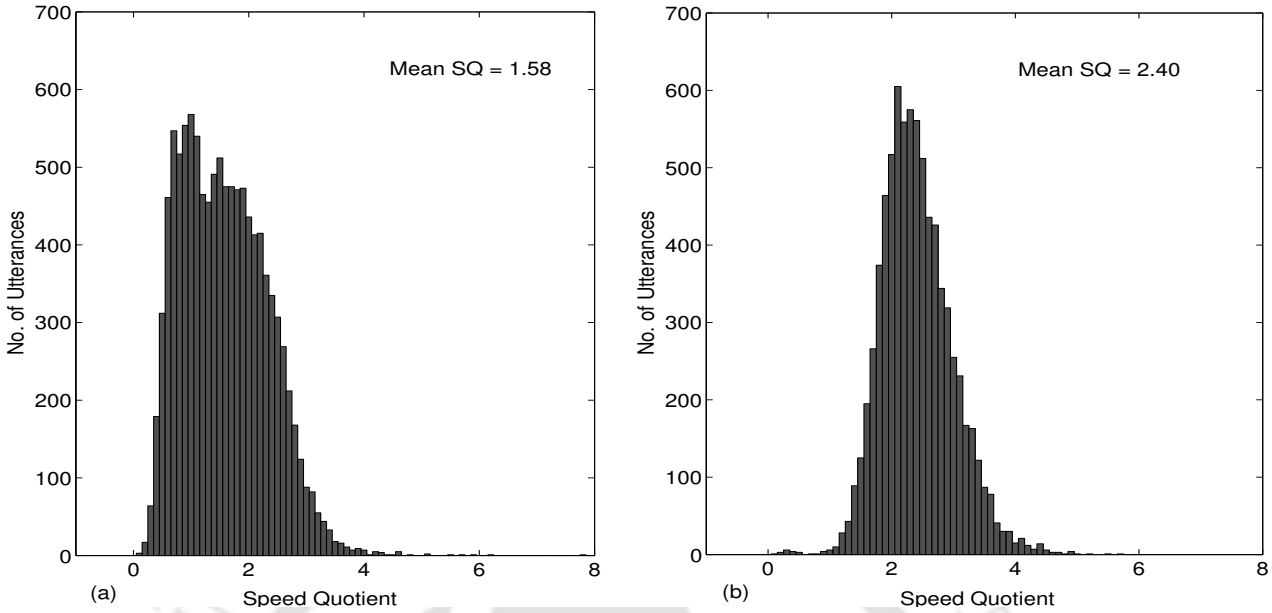


Figure 3.9: Distribution of average SQ of the original signals of (a) adults' training set ADtr (b) children's test set CHts1.

in reducing the mismatch in formant frequencies for two speakers can be understood by observing the original smoothed Mel spectrum and those obtained after warping the frequency axis of the speech spectrum by factors 0.88 and 1.12 for a central steady-state portion of vowel /IY/ having pitch value of around 100 Hz as shown in Figure 3.10.

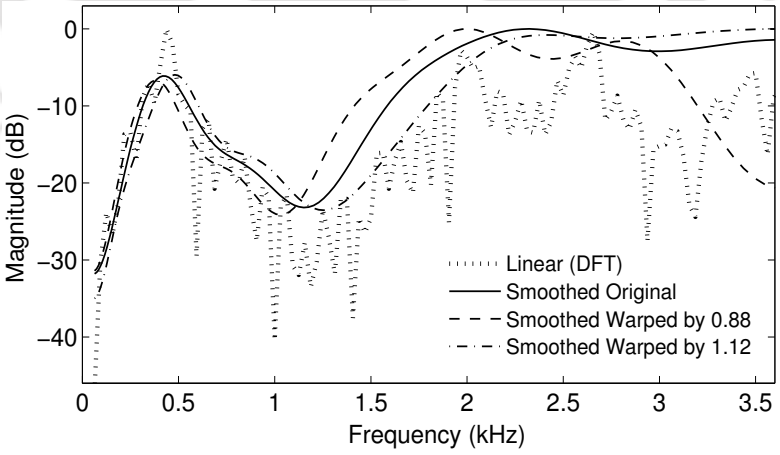


Figure 3.10: Plots showing original and frequency warped smoothed Mel spectra for vowel /IY/ having pitch value of around 100 Hz.

3.3 Relative Significance of Various Acoustic Sources of Mismatch for Children's ASR

In the previous section, the effect of variations in various acoustic correlates of speech has been noted on MFCC features and on the HMM-based ASR models. Also, the degree of differences in those acoustic parameters existing between the adults' training and the children's test speech data that is used in this work has been observed. Following those, in this section, the relative significance of each of those acoustic sources of mismatches is explored for children's speech recognition on the adults' speech trained models. The acoustic correlates explored in this study are the pitch, the speaking rate, the formant frequencies and the glottal flow parameters (OQ, RQ and SQ). First, the study is performed on a connected digit (limited vocabulary) recognition task and later, it is repeated for comparatively a large vocabulary continuous speech recognition task. The databases used in this section for various recognition experiments are described in Section 2.2.

To explore the relative significance of various acoustic correlates of speech for children's ASR on the adults' speech trained acoustic models, the differences in those acoustic parameters of children's speech signals with respect to the adults' training speech data set are explicitly normalized. To determine the optimal value to which each of the acoustic correlates is to be transformed to, a ML grid search is used. For instance, the optimal value of an acoustic correlate, say $\hat{\beta}$, given its various transformed values within the valid range, is estimated as:

$$\hat{\beta} = \arg\max_{\beta} P(X_i^{\beta}|\lambda_{ad}, W_i) \tag{3.2}$$

where, X_i^{β} represents the feature corresponding to a particular value β of an acoustic correlate for the i^{th} children's test speech utterance, λ_{ad} represents the adults' speech trained recognition model and W_i represents the transcription of the i^{th} utterance. The W_i is determined by doing a first-pass recognition using original features (i.e., without transformation of the acoustic correlate) with respect to the adults' speech trained models. The use of the erroneous first-pass transcription, W_i, is definitely a compromise with respect to using the true (reference) transcription. However, the use of the true transcription for the normalization of the test data would not be valid and feasible in practical cases.

The average pitch of the signal, the signal duration and the glottal flow parameters of the signal are transformed explicitly in the signal domain prior to feature computation whereas the formant frequencies are modified in the feature domain as described in Chapter 2. In the following subsections,

the experimental conditions and the results obtained after explicit normalization of each of these acoustic correlates independently on both the connected digit recognition and the continuous speech recognition tasks are described in detail.

3.3.1 Connected Digit Recognition Task

The adults' speech trained models used in this study for the digit recognition task are developed using the adults' training set ADtr derived from the TIDIGITS corpus. The recognition performance (in WER) for the adults' test set ADts and the children's test set CHts1 are 0.43% and 11.37%, respectively. The recognition results obtained for children's speech recognition after explicit normalization of the various acoustic correlates of children's speech are given in the following subsections.

3.3.1.1 Pitch

For explicit pitch normalization of the children's test set CHts1, all children's test speech signals are transformed to seven different pitch values ranging from 70 Hz to 250 Hz in steps of 30 Hz using PSTS method as described in Section 2.4.1.1. Such pitch range has been chosen based on the distribution of average pitch of the signals of training set ADtr as shown in Figure 3.3(a). The average pitch of a speech signal is estimated using the ESPS tool available in the Wavesurfer software package [88]. The quality of pitch transformation and signal reconstruction by PSTS method can be understood by observing the spectrograms for a speech signal before and after its explicit pitch transformation. Figure 3.11(a) and Figure 3.12(a) show the spectrograms and the contours of the first four formant frequencies of two voiced portions corresponding two different words extracted from a speech signal of the TIDIGITS database having an average pitch value of 200 Hz viz., 'Three' and 'Four', respectively. The spectrograms and the contours of the first four formant frequencies of the two words 'Three' and 'Four' extracted from the speech signal after explicit transformation of its average pitch value to 130 Hz are shown in Figure 3.11(b) and Figure 3.12(b), respectively. It is noted that after pitch transformation by PSTS method only the pitch harmonics have significantly shifted to lower frequencies while the overall variation in the formant structure of the signal is much smaller. This can be also validated by observing the mean values of the first four formant frequencies for the two words before and after explicit pitch transformation of the signal given in Table 3.2.

For determining the appropriate transformations of pitch values for each of the children's test speech signals, a ML grid search is done among the original signal and its seven pitch transformed

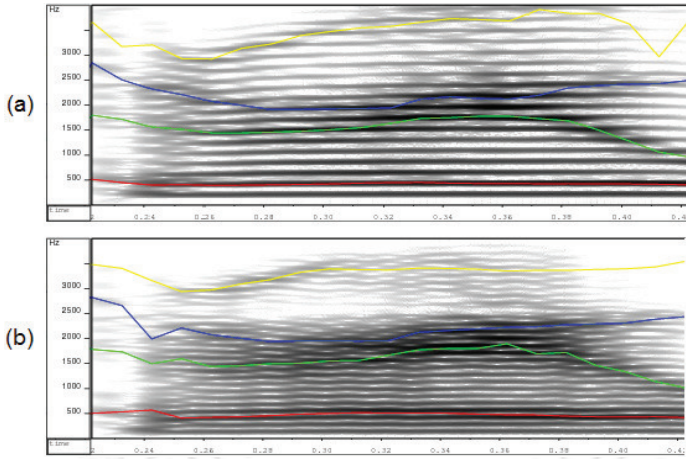


Figure 3.11: Spectrogram of a voiced portion corresponding to word 'Three' extracted from a speech utterance before and after explicit transformation of its average pitch value from 200 Hz to 130 Hz by PSTS method (a) Original 200 Hz (b) Pitch transformed to 130 Hz. The red, green, blue and yellow line plots are the contours of the first, second, third and fourth formants, respectively.

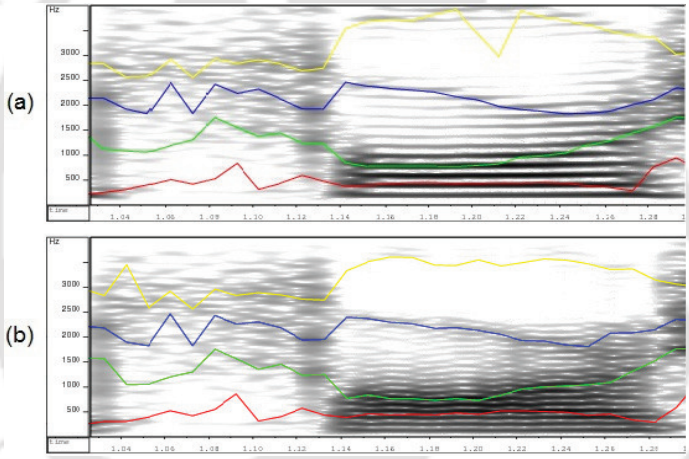


Figure 3.12: Spectrogram of a voiced portion corresponding to word 'Four' extracted from a speech utterance before and after explicit transformation of its average pitch value from 200 Hz to 130 Hz by PSTS method (a) Original 200 Hz (b) Pitch transformed to 130 Hz. The red, green, blue and yellow line plots are the contours of the first, second, third and fourth formants, respectively.

versions similar to the one commonly used for ML-based speaker normalization [55]. The recognition performances of the children's test set CHts1 with and without explicit pitch normalization are given in Table 3.3 along with their pitch group-wise breakup. To study the effect of the pitch mismatch of the children's test data with respect to the pitch range of the adults' training data, the pitch groups of $F_o <$ 250 Hz, 250 Hz $\leq F_o <$ 300 Hz and $F_o >$ 300 Hz are chosen for the children's test set. The pitch group $F_o <$ 250 Hz matches with the pitch range of the adults' training speech data. It is noted that explicit pitch normalization of children's speech results in 15% relative improvement in the ASR performance

Table 3.2: Mean values of the first four formant frequencies for two different words extracted from a speech signal of the TIDIGITS database before and after its explicit pitch transformation by the PSTS method.

Condition	Words							
	'Three'				'Four'			
	Formants (in Hz)				Formants (in Hz)			
	First	Second	Third	Fourth	First	Second	Third	Fourth
Original (F_o = 200 Hz)	426	1529	2178	3459	723	1262	2043	2907
Explicitly Pitch Transformed to 130 Hz	475	1558	2167	3311	693	1274	2324	3189

for CHts1 test set. For test signals having average pitch value before transformation in the range of $F_o < 250$ Hz, 250 Hz $\leq F_o < 300$ Hz and $F_o > 300$ Hz, a relative improvement of about 8%, 19% and 23% is obtained after explicit pitch normalization, respectively. Thus, consistent improvements are noted for the different pitch groups i.e., higher pitch groups have greater improvements.

The improvements obtained with explicit pitch normalization can be further understood by observing the pitch distribution of the adults' training set ADtr and the children's test set CHts1 before and after explicit pitch normalization as shown in Figure 3.3(a), Figure 3.3(b) and Figure 3.13, respectively. It is noted that the mean of the pitch distribution of the children's test set CHts1 has shifted towards that of the adults' training set ADtr after ML-based explicit pitch normalization. Also, on comparing the means of the pitch distributions for different pitch groups of children's test

Table 3.3: Performance for children's test set CHts1 (with breakup for different pitch groups based on original average pitch values) with and without explicit pitch normalization. The quantity in parentheses shows the number of utterances in that group. The 95% confidence interval for the performance for CHts1 data set is ±0.39 [for $F_o < 250$ Hz, 250 Hz $\leq F_o < 300$ Hz and $F_o > 300$ Hz pitch groups ±0.39, ±0.79 and ±3.37, respectively].

Condition	% WER			
	All F_o Values (7,772)	$F_o <$ 250 Hz (5,224)	250 Hz $\leq F_o <$ 300 Hz (2,346)	$F_o \geq$ 300 Hz (202)
Baseline	11.37	6.54	17.47	39.03
Pitch Norm.	9.64	6.02	14.24	30.11

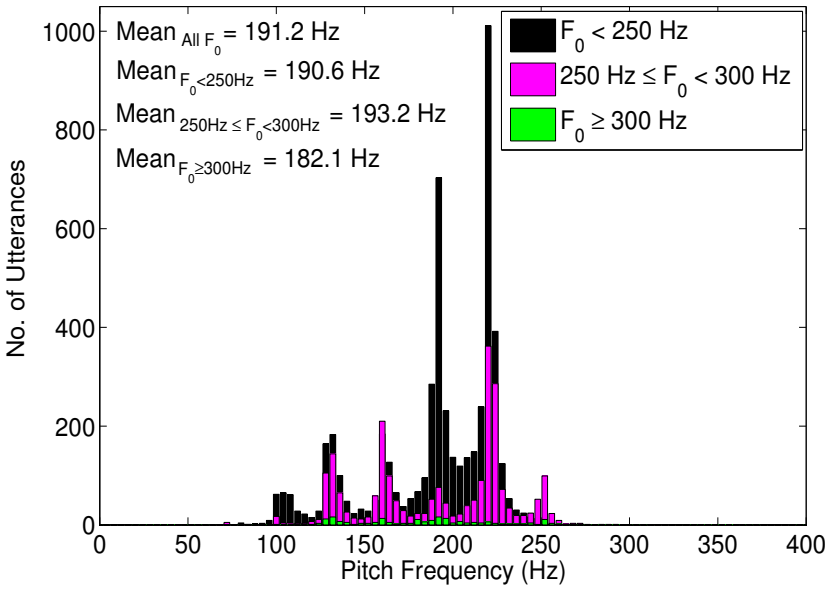


Figure 3.13: Distributions of the average pitch of the signals of the children's test set CHts1 after ML-based explicit pitch normalization for all the three pitch groups (as defined in Figure 3.3(b) based on their original pitch values) plotted separately.

set CHts1 before and after explicit pitch normalization, it is noted that the shift in the mean of the pitch distribution is more for higher pitch group signals.

These improvements in the recognition performance of children's speech with explicit pitch normalization could be attributed to the reduction in the variances of the higher order MFCCs of the children's test speech signals with reduction in the average pitch values of the signals. As a result, the MD for the children's test signals with respect to the adults' speech trained models is reduced leading to better classification performance of the models for the children's test speech as discussed in Section 3.2.1. This is further supported by the improvements obtained in the children's ASR performance on adults' speech trained models on applying variance normalization to 39-dimensional MFCC features reported in [78].

3.3.1.2 Speaking Rate

For explicit normalization of the speaking rate of the children's test set CHts1 according to that of the adults' speech trained models, the duration of the signals is reduced by factors ranging from 0.6 to 1.0 in steps of 0.05, thereby increasing the speaking rate of the signals by factors ranging from 1.0 to 1.65. The choice of such duration transformation factors is based on the distribution of the speaking rate of the signals belonging to the adults' training set as shown in Figure 3.5(a). Note that

Table 3.4: Performance for children's test set CHts1 with and without explicit normalization of different acoustic correlates of speech. The 95% confidence interval for the performance for CHts1 data set is ±0.39.

Condition	% WER
Baseline	11.37
Pitch Norm.	9.64
Speaking Rate Norm.	10.31
Open Quotient Norm.	11.32
Return Quotient Norm.	11.28
Speed Quotient Norm.	11.01
Formant Freq. Norm.	2.95

the average speaking rate of the signals on the digit recognition task is measured as the number of syllables per second computed as the ratio of the number of syllables in an utterance to the total length of the utterance. Each of the 11 digits constituting the training and test set utterances used in this work, comprise of only 1 syllable except for the digits 'zero' and 'seven' which contain 2 syllables. For determining the appropriate transformations of duration of each of the children's test speech signals, a ML grid search is done similar to the one commonly used for ML-based speaker normalization [55].

The recognition performances of the children's test set CHts1 with and without explicit speaking rate normalization are given in Table 3.4. It is noted that explicit speaking rate normalization results in a significant 9% relative improvement over the baseline performance. This improvement is consistent with the results reported in literature obtained with explicit speaking rate normalization for ASR of signals of mismatched speaking rate [59, 113]. The distributions of the speaking rate of the adults' training set ADtr and the children's test set CHts1 before and after explicit speaking rate normalization are shown in Figure 3.5(a), Figure 3.5(b) and Figure 3.14, respectively. On comparing these distributions, it is noted that the mean speaking rate of the children's test set CHts1 has been transformed towards that of the adults' training set ADtr after explicit speaking rate normalization.

Further, on comparing the log likelihoods of the signals from the children's test set CHts1 on the adults' speech trained models before and after explicit speaking rate normalization, shown in Figure 3.15, it is noted that the likelihood of all of the children's speech utterances has increased after explicit normalization of their speaking rate. This verifies the reduction in the earlier hypothesized mismatch in the duration modeling of the children's test set with respect to the adults' speech trained models.

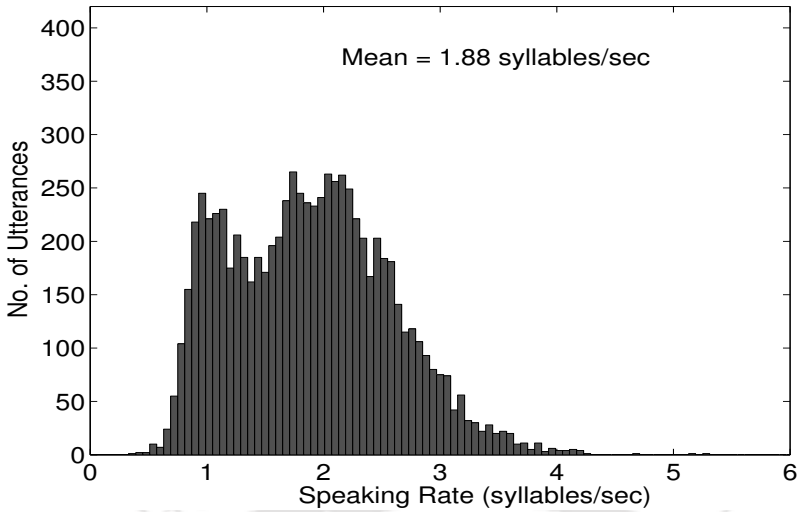


Figure 3.14: Distribution of average speaking rate of signals of the children's test set CHts1 after explicit speaking rate normalization.

3.3.1.3 Glottal Flow Parameters

For explicitly normalizing the variations in the glottal flow parameters of children's test set CHts1 with respect to those of the adults' training set ADtr, a ML grid search is performed to determine the appropriate transformation for each of the three glottal flow parameters of each signal. The ML grid search is done among the transformed versions of the signal with OQ, RQ and SQ of the signal being modified by factors ranging from 0.55 to 1.0, 0.35 to 1.0 and 0.45 to 1.0 each in steps of 0.05,

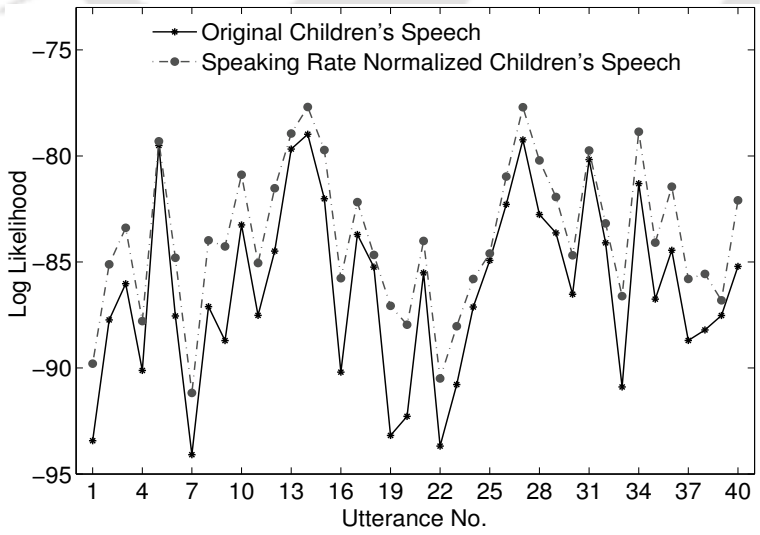


Figure 3.15: Log likelihood distribution of few utterances from the children's test set CHts1 before and after explicit speaking rate normalization on models trained with adults' training set ADtr.

respectively. The choice of such transformation factors for OQ, RQ and SQ modification is supported by the studies done in literature which report that children's speech have higher OQ, RQ and SQ values than those of the adults' speech [33, 34, 114]. The ASR performances for the children's test set CHts1 with and without explicit normalization of the three glottal flow parameters are given in Table 3.4. It is noted that no significant improvement is obtained over the baseline with explicit normalization of any of these glottal flow parameters. Although, the glottal flow parameters have been found to be of significance in case of one-to-one voice transformation [43, 115] but in case of ASR, where the acoustic models are trained using data from a large number of speakers, enough variations in the glottal flow parameters are captured within the training data itself. As a result, a very little mismatch exists due to differences in the glottal flow parameters between the adults' training and the children's test data.

The age group-wise distribution of the ML-based transformation factors chosen for explicit normalization of OQ, RQ and SQ of the children's test speech signals with respect to the adults' speech trained models are shown in Figure 3.16, Figure 3.17(a) and Figure 3.17(b), respectively. It is noted that for explicit normalization of each of these glottal flow parameters, majority of the signals have opted for no transformation across all age groups. Also, it is worth noting that all transformation factors have been chosen by the signals of all age groups in similar proportion. Thus, there seems to be very little correlation between the age and the glottal flow parameters (OQ, RQ, SQ).

3.3.1.4 Formant Frequencies

The variations in formant frequencies of the adults' and the children's speech occur due to differences in their vocal tract lengths which is usually modeled as a constant scaling of the resonant peaks in the spectral domain. For explicitly normalizing the variations in the formant frequencies of the signals of the children's test set CHts1, a ML grid search is performed among features warped by 13 equally spaced warping factors ranging from 0.88 to 1.12 in steps of 0.02 for each signal as described in Section 2.4.2. The ASR performances of the children's test set CHts1 with and without VTLN are given in Table 3.4. It is noted that VTLN results in a large relative improvement of 74% over baseline for children's ASR performance.

3.3.2 Continuous Speech Recognition Task

On observing the improvements obtained in the children's ASR performance on a connected digit recognition task by explicit normalization of various acoustic correlates of children's speech given in

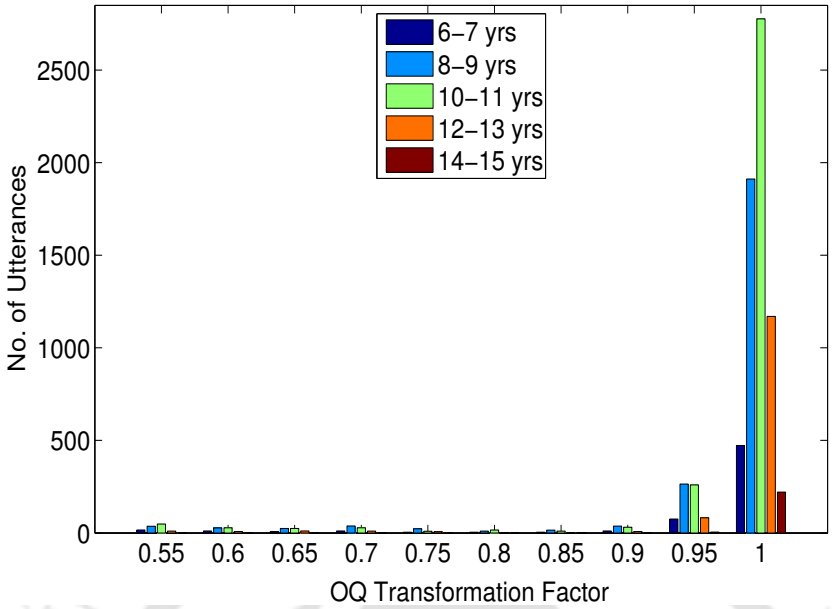


Figure 3.16: Age group-wise distribution of optimal OQ transformation factors chosen for the signals of the children's test set CHts1.

Table 3.4, it is noted that only the formant frequencies, the pitch and the speaking rate affect the children's ASR performance significantly. Thus, the effectiveness of explicit normalization of these three acoustic sources only in improving the children's ASR performance is further validated and

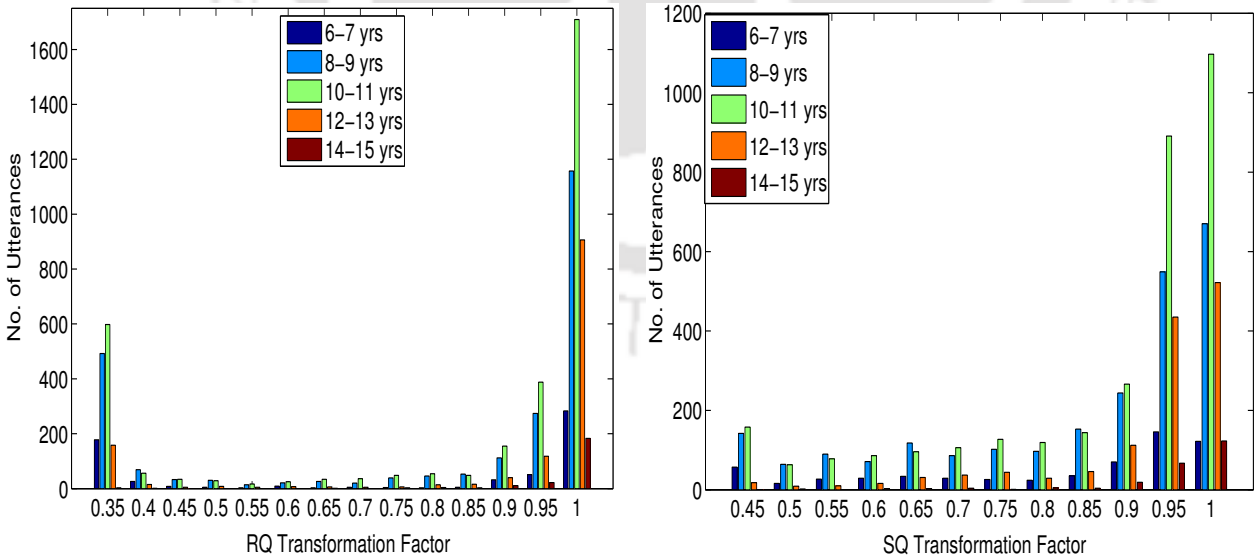


Figure 3.17: Age group-wise distribution of optimal transformation factors chosen for the signals of the children's test set CHts1 for (a) RQ (b) SQ.

Table 3.5: Performance for children's test set PFts with and without explicit normalization of different acoustic correlates of speech. The 95% confidence interval for the performance for PFts data set is ±1.37.

% WER			
Baseline	Explicit Normalization		
	Formant Frequencies	Pitch	Speaking Rate
56.34	26.78	53.72	54.27

quantified on comparatively a large vocabulary continuous speech recognition task. For continuous speech recognition task, the adults' speech trained models are developed using the adults' training set CAMtr. The recognition performance (in WER) for the adults' test set CAMts and the children's test set PFts are 9.92% and 56.34%, respectively.

For VTLN, frequency warping factors ranging from 0.88 to 1.12 in steps of 0.02 are used for each signal. For explicit pitch normalization, the average pitch of each test signal of PFts test set is transformed to seven different pitch values ranging from 80-260 Hz in steps of 30 Hz. For explicit speaking rate normalization, the duration of each test signal of PFts test set is transformed by factors ranging from 0.5-1.0 in steps of 0.05. This range of values for pitch and duration modification was chosen based on the average pitch and speaking rate distribution of the adults' speech training set CAMtr. Note that the average speaking rate of the signals is measured as the number of phones per second. Given the various VTL normalized, pitch transformed, duration transformed versions within the specified range, the optimal values for each test speech signal in each case are estimated using the ML grid search in a similar manner as described above.

The recognition performances of the PFts test set with and without normalization of the formant frequencies, the pitch and the speaking rate are given in Table 3.5. It is noted that after explicit normalization of the formant frequencies, the pitch and the speaking rate the recognition performance for children's speech improves over baseline relatively by 52.5%, 4.7% and 3.7%, respectively. Therefore, besides formant frequencies, the pitch is found to be the other major source of acoustic mismatch which significantly affects the children's ASR performance and is found to give consistent improvement in the ASR performance after explicit normalization on both the connected digit and the continuous speech recognition tasks.

3.4 Combining VTLN and Explicit Acoustic Normalization with Model Adaptation

In previous section, it is noted that independent normalization of various acoustic sources of mismatch significantly improves the children's speech recognition performance. In literature, the model adaptation techniques viz., MLLR and CMLLR have also been reported for reducing or compensating for acoustic variations induced by differences in the characteristics of the training and the test speakers. However, these adaptation techniques employ linear transformations with no assumption about the specific nature of acoustic mismatch being addressed. Since for all different sources of acoustic mismatch between the adults' and the children's speech the linear transformation model may not be appropriate, it would be better to compensate them explicitly. Motivated by this, the efficacy of explicit normalization of different acoustic correlates is explored in conjunction with MLLR and CMLLR model adaptation techniques.

In order to explore the independent effect of explicit normalization of different acoustic correlates on the means and the variances of the models, transformations of both means and variances of the models are also learnt independently in addition to their combination. In MLLR adaptation method, when the affine transformations are applied to the mean parameters of the models only it is referred to as 'MLLR-MEAN' and when the linear transformations are applied to the covariance parameters of the models only it is referred to as 'MLLR-COV'. The recognition results for the original, the explicitly pitch normalized and the explicitly speaking rate normalized CHts1 and PFts test sets on the corresponding adults' speech trained models with and without various model adaptation techniques viz., MLLR-MEAN, MLLR-COV and CMLLR are given in Table 3.6 and Table 3.7, respectively. For sake of comparison, the recognition results for the CHts1 and PFts test sets with VTLN both with and without various model adaptation techniques viz., MLLR-MEAN, MLLR-COV and CMLLR for the original children's speech are also given in Table 3.6 and Table 3.7, respectively. It is to note that all model adaptations and explicit acoustic normalization are applied only to the children's test sets in the manner as described in Chapter 2.

From Table 3.6, it is noted that for original children's test set CHts1 relative improvements of 46.9%, 41.2% and 51.3% are obtained over baseline with MLLR-MEAN, MLLR-COV and CMLLR model adaptations, respectively. Consistent relative improvements of 46.5%, 35.8% and 47.2% are obtained for the explicitly pitch normalized test set CHts1 with MLLR-MEAN, MLLR-COV and

Table 3.6: Performance for children's test set CHts1 with and without explicit pitch and speaking rate normalization, VTLN and model adaptation. The numbers given in parentheses are the relative improvements (in %) obtained with respect to their corresponding baseline.

Condition	% WER			
	Baseline	MLLR-MEAN	MLLR-COV	CMLLR
Original	11.37	6.04 (46.9)	6.68 (41.2)	5.54 (51.3)
Formant Freq. Norm.	2.95	1.72 (41.7)	1.79 (39.3)	1.51 (48.8)
Pitch Norm.	9.64	5.16 (46.5)	6.19 (35.8)	5.09 (47.2)
Speaking Rate Norm.	10.31	5.32 (48.4)	6.18 (40.1)	4.84 (53.1)

CMLLR, respectively. Consistent relative improvements of 48.4%, 40.1% and 53.1% are obtained for the explicitly speaking rate normalized children's test set CHts1 with MLLR-MEAN, MLLR-COV and CMLLR adaptations, respectively. Similarly, from Table 3.7, it is noted that for original PFts data set relative improvements of 28.3%, 23.2% and 32.1% are obtained over baseline with MLLR-MEAN, MLLR-COV and CMLLR model adaptations, respectively. Consistent relative improvements of 31.3%, 19.8% and 29.4% are obtained for the explicitly pitch normalized children's test set PFts with MLLR-MEAN, MLLR-COV and CMLLR adaptations, respectively. Consistent relative improvements of 29.4%, 22.4% and 31.4% are obtained for the explicitly speaking rate normalized children's

Table 3.7: Performance for children's test set PFts with and without explicit pitch and speaking rate normalization, VTLN and model adaptation. The numbers given in parentheses are the relative improvements (in %) obtained with respect to their corresponding baseline.

Condition	% WER			
	Baseline	MLLR-MEAN	MLLR-COV	CMLLR
Original	56.34	40.38 (28.3)	43.26 (23.2)	38.25 (32.1)
Formant Freq. Norm.	26.78	20.66 (22.9)	19.42 (27.5)	18.63 (30.4)
Pitch Norm.	53.72	36.93 (31.3)	43.10 (19.8)	37.91 (29.4)
Speaking Rate Norm.	54.27	38.31 (29.4)	42.14 (22.4)	37.24 (31.4)

test set PFts with MLLR-MEAN, MLLR-COV and CMLLR adaptations, respectively. From these results, it is noted that CMLLR gives the best ASR performance with maximum improvement over the corresponding baseline among all the three model adaptation methods which is also consistent with the literature. This observation is made for the original as well as the explicitly pitch normalized and the speaking rate normalized children's speech on both connected digit and continuous speech recognition tasks. Similar observation is also made when the model adaptations are done over VTLN for original children's speech. Thus, the improvements obtained in the children's ASR performance with the existing model adaptation techniques are consistent even after explicit normalization of various studied acoustic correlates for children's speech.

Observing the relative improvements obtained with each of the model adaptation methods before and after explicit pitch normalization, it is to note that though after explicit pitch normalization the relative improvements obtained with each of the model adaptation techniques are reduced, greater reduction in the relative improvement is obtained for MLLR-COV and CMLLR than for MLLR-MEAN on both tasks. This is attributed to the significant reduction in the variances of the features after explicit pitch normalization as observed in Section 3.2.1 which thus, reduces the scope of improvement with the model adaptations which transform the variances of the models. However, it is worth noting here that further relative improvements of 14.6%, 7.3% and 8.1% are obtained on additionally doing explicit pitch normalization for the children's test set CHts1 over the performances obtained with MLLR-MEAN, MLLR-COV and CMLLR, respectively. Similarly, for children's test set PFts further relative improvements of 8.5%, 0.4% and 0.9% are obtained on additionally doing explicit pitch normalization over the performances obtained with MLLR-MEAN, MLLR-COV and CMLLR, respectively.

Thus, additional improvements are obtained with explicit pitch normalization over those obtained with existing linear transformation based model adaptation techniques which are attributed to the fact that unlike in latter, with explicit pitch normalization the transformations are not constrained to be linear.

3.5 Summary

In this chapter, the effect of differences in various acoustic sources of mismatch between the adults' and the children's speech on MFCC features and HMM-based ASR models is explored. The

various acoustic correlates that are studied in this work are the pitch, the speaking rate, the formant frequencies and the glottal flow parameters (OQ, RQ, SQ). In addition to this, the relative significance of each of these acoustic correlates is explored for children's ASR on adults' speech trained models on both the connected digit recognition and the continuous speech recognition tasks. The salient observations made in this chapter are:

- With increase in the pitch of the signals, the variances of the higher order coefficients of the 13-dimensional default base MFCC features are noted to increase significantly.
- The rate of speech of training data significantly affects the state-transition probabilities of the HMM-based acoustic models. The adults' speech trained models are noted to have lower self-loop transition probabilities than those of the children's speech trained models.
- Differences in the formant frequencies of adults' and children's speech appear as systematic scaling of the smoothed Mel spectra corresponding to their MFCC features.
- It is found that, besides formant frequencies, the pitch is the other major source of acoustic mismatch which significantly degrades the children's ASR performance on adults' speech trained models.
- No significant effect of variations in glottal flow parameters is noted on the children's ASR performance on adults' speech trained models.
- Consistent and significant improvement is obtained in the children's ASR performance on both the connected digit recognition and the continuous speech recognition tasks after explicit pitch normalization of children's speech.
- The improvement in children's ASR performance with explicit pitch normalization is also found to be additive to those obtained with the existing model adaptation techniques viz., MLLR and CMLLR.

Observing the increase in the variances of higher order coefficients of MFCC features with increase in the pitch of the signals, in the next chapter, the cause and the nature of the effect of pitch on MFCC features is explored in detail.

4

Effect of Pitch on MFCC Features

Contents

4.1 Introduction 62
4.2 Effect of Uniform and Non-Uniform Filterbank on Pitch Harmonicity . 63
4.3 Effect of Pitch-dependent Distortions on MFCCs 71
4.4 Summary 74

4.1 Introduction

In ASR, the objective is to recognize 'what is spoken' rather than 'who has spoken'. As a result, all features used for ASR aim at capturing only the vocal tract filter characterizing the phone and discarding the speaker-dependent information like pitch harmonics [116,117]. It is well known that the relevant information about the vocal tract filter is mainly encoded in the envelope of the short-time speech spectrum [118]. One of the traditional features used in ASR employed LP analysis to capture the smoothed speech spectral envelope. The resulting LPC were converted to cepstral coefficients for employing the Euclidean distance measure and to decorrelate the feature coefficients [119]. In addition to that, the smoothed spectrum is also derived by employing a filterbank either uniform or non-uniform. To get the advantages of the cepstral representation, the filterbank energies are also further converted to cepstral coefficients. The cepstrum allows the deconvolution of the periodic voiced excitation signal from the effects of the vocal tract filter. Based on this, Noll in his pioneering work [120], proposed a cepstral analysis based pitch estimation approach. Motivated by that, the cepstral smoothing using a low-time lifter was explored by Schafer and Rabiner [121] for formant analysis. Following those studies, all current cepstral features often use the low-time liftering in quefrency domain to capture a smoothed spectral envelope devoid of the pitch-related information.

On the contrary, in [99], it has been reported that a limited amount of pitch information is retained in MFCC features. Following it, the possibility of predicting the pitch of a signal and thus, the feasibility of reconstructing a speech signal from its MFCC features has been explored. Also, the improvement in the phone classification performance with pitch-dependent normalization of Mel cepstrum has been reported in [100]. Similarly, in the previous chapter, we have also noted a significant effect of pitch on MFCC features. Significant increase in the variances of the higher order MFCCs has been observed with increase in the pitch of the speech signals. However, despite the filterbank based spectral smoothing and cepstral truncation using a low-time lifter in quefrency domain, how the effect of the pitch manifests in the MFCC features is not very obvious. Though in [116], Hunt has briefly mentioned that MFCC features may get effected by pitch variations due to the use of the Mel filterbank, but otherwise this issue has not been explained in detail in the literature. Motivated by these, in this chapter, we attempt to explore the cause and the nature of the effect of pitch on MFCC features.

The rest of the chapter is organized as follows: In Section 4.2, the roles of filterbank and cepstral

truncation are studied in removing the pitch-related information from the uniform filterbank based and the non-uniform Mel filterbank based spectra and their corresponding cepstra. The nature and the effect of the pitch-dependent distortions appearing in the Mel spectral envelope on MFCC features are explored in Section 4.3. Finally, the observations made in this chapter are summarized in Section 4.4. The vowel speech data extracted from TIMIT corpus is used for the analysis in this chapter. To be consistent with the other experiments reported in this thesis, the speech data used in this analysis is also sampled to 8 kHz.

4.2 Effect of Uniform and Non-Uniform Filterbank on Pitch Harmonicity

In this section, it is highlighted that the cepstral truncation using a low-time lifter is able to remove the pitch-related information from a speech spectrum obtained using uniform filterbank based spectral analysis only. To demonstrate this, the uniform filterbank based and the non-uniform filterbank based spectra and their corresponding cepstra for central steady-state portions of vowels from low and high pitch signals are considered. The average pitch of the signals is determined using the ESPS tool as described in Section 2.2.

The uniform filterbank based spectrum is obtained by employing a uniform 30-channel triangular filterbank while the non-uniform filterbank based spectrum is obtained by employing a 30-channel triangular Mel filterbank (as per HTK implementation) on the 128-point linear DFT spectrum. The 30-channel filterbanks are used so that the pitch harmonicity, if present, could be observed in the cepstra of the high pitch signals. The speech analysis for estimating the 128-point linear DFT spectrum is done following the configuration described in Section 2.2. The corresponding cepstra for both the uniform filterbank based spectrum and the Mel spectrum are computed by taking the discrete cosine transform (DCT) of the log-compressed spectra.

4.2.1 Uniform Filterbank based Spectral Analysis

In this section, we demonstrate the roles of filterbank and cepstral truncation in removing the pitch-related information from the uniform filterbank based speech spectra and their corresponding cepstra for different pitch signals. The purpose of this demonstration is to provide a contrast to the non-uniform filterbank based spectral analysis study discussed in the next section.

For sake of reference, the linear DFT spectra and their corresponding cepstra are first shown for

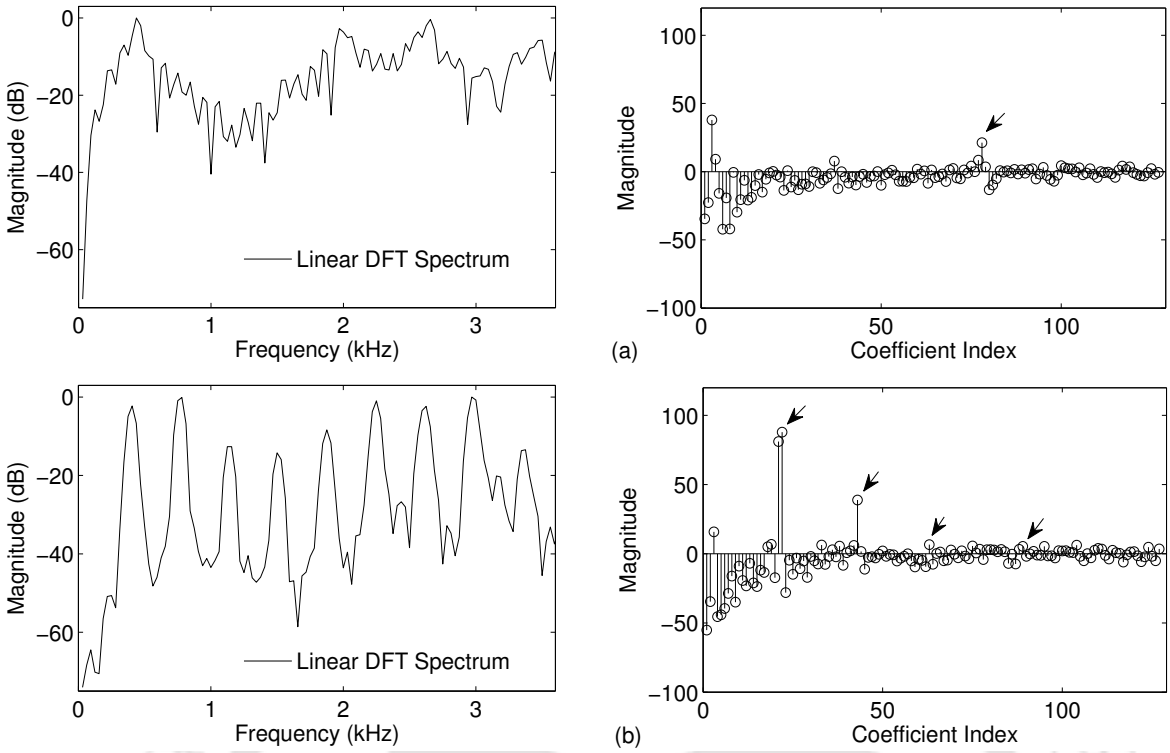


Figure 4.1: Plots of the 128-point linear DFT spectra (left panel) and their corresponding cepstra (right panel) for vowel /IY/ having pitch values of around (a) 100 Hz (typical value for male adults' speech) (b) 300 Hz (typical value for children's speech). The peaks in the cepstra corresponding to the pitch harmonics are marked with arrows. Note that for clarity the plots are shown excluding the C_0 coefficient.

different pitch signals. The plots of the 128-point linear DFT spectra and their corresponding 128-point cepstra for vowel /IY/ having pitch values of around 100 Hz ('low') and 300 Hz ('high') are shown in Figure 4.1. It is noted that, as expected, the cepstra derived from the linear DFT spectra exhibit peaks at the multiples of the pitch period of the corresponding speech frame in case of both low and high pitch vowel frames. These periodic peaks appearing in the cepstra closely correspond to the pitch harmonics observed in their corresponding linear DFT spectra.

In order to understand the effect of an explicit filterbank applied to a linear DFT spectrum on the spectral analysis, we first explore a uniform filterbank applied to the linear DFT spectra of different pitch signals. The plots of the 30-point uniform filterbank based spectra and their corresponding 30-point cepstra for vowel /IY/ having pitch values of around 100 Hz ('low') and 300 Hz ('high') are shown in Figure 4.2. It is to note that for low pitch vowel frame the uniform filterbank based spectrum is completely smoothed out and correspondingly no pitch harmonicity is noted in its cepstrum. However, the pitch harmonicity is clearly observed in the uniform filterbank based spectrum for high pitch vowel

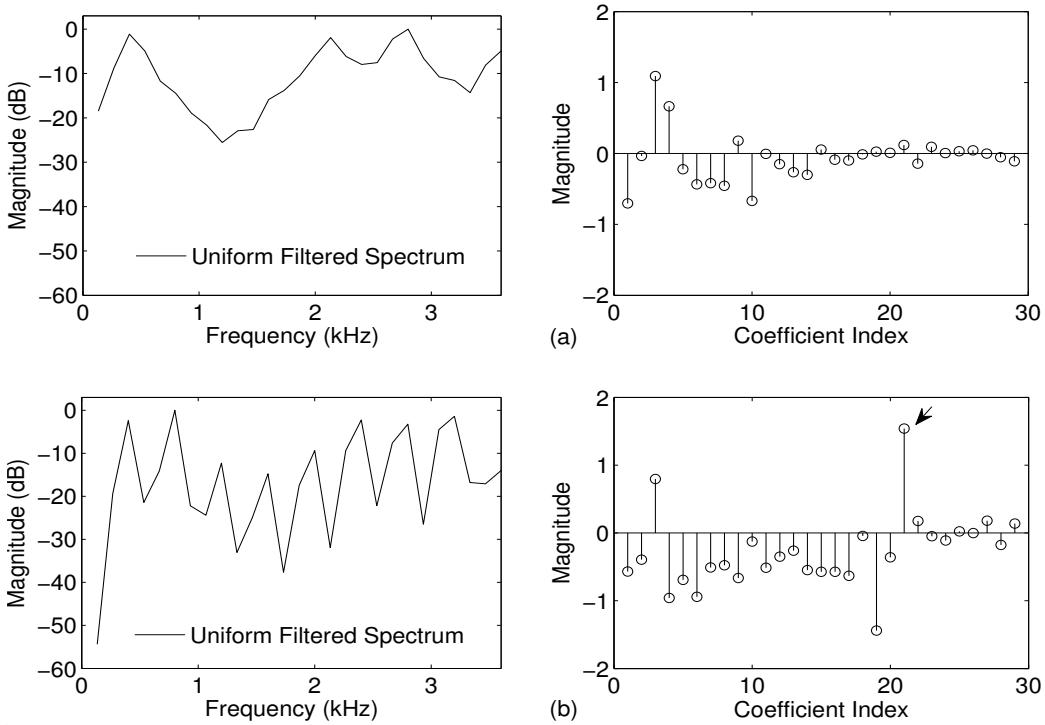


Figure 4.2: Plots of the 30-point uniform filterbank based spectra (left panel) and their corresponding cepstra (right panel) for vowel /IY/ having pitch values of around (a) 100 Hz (b) 300 Hz. The arrow in the cepstrum of the 300 Hz pitch signal shows the location of the first pitch harmonic.

frame. As a result, the uniform filterbank based cepstrum also exhibits the pitch harmonicity with only observable peak corresponding to the first pitch harmonic occurring at the identical location as that observed in the cepstrum derived from the linear DFT spectrum for high pitch vowel frame. In case of low pitch vowel frame, the bandwidths of the filters are larger than the separation between the pitch harmonics in the linear DFT spectrum while they are smaller in case of high pitch vowel frame. This results in smoothing of the pitch harmonics in the linear DFT spectrum of low pitch vowel frame while they are preserved in case of high pitch vowel frame. Thus, on account of uniform filtering of the speech spectrum, the pitch harmonics in the resulting spectrum are either smoothed out or their regularity is maintained. Correspondingly, the cepstrum also exhibits either no or similar pitch harmonicity. Therefore, by truncating the uniform filterbank based cepstrum with a low-time rectangular lifter having length lesser than the pitch period of the signal, the effect of pitch can be removed from the cepstrum.

4.2.2 Non-Uniform Filterbank based Spectral Analysis

To explore the effect of applying a non-uniform filterbank to a linear DFT spectrum for spectral analysis, in this section we compare the smoothed Mel spectra and their corresponding cepstra obtained using the non-uniform Mel filterbank for different pitch signals. The plots of the 30-point Mel spectra and their corresponding 30-point cepstra for vowel /IY/ having pitch values of around 100 Hz ('low') and 300 Hz ('high') are shown in Figure 4.3. It is noted that in case of low pitch signal, the Mel spectral envelope closely follows the spectral envelope of the linear DFT spectrum with no significant pitch harmonicity. Like Mel spectrum, the corresponding cepstrum for the low pitch signal also does not appear to contain any pitch harmonicity. On the other hand, in case of the high pitch signal, some distortions are noted in the Mel spectral envelope which appear to correspond to the pitch harmonics present in the linear DFT spectrum of the high pitch signal. However, no such pitch-dependent harmonicity is observed in the Mel cepstrum of the high pitch signal. Similar behavior is noted in the default 21-channel Mel spectra and their corresponding cepstra for vowel /IY/ having pitch values of around 100 Hz ('low') and 300 Hz ('high') also as shown in Figure 4.4.

For ASR, usually 13-D truncated base Mel cepstral features i.e., MFCC features ($C_0 - C_{12}$) derived

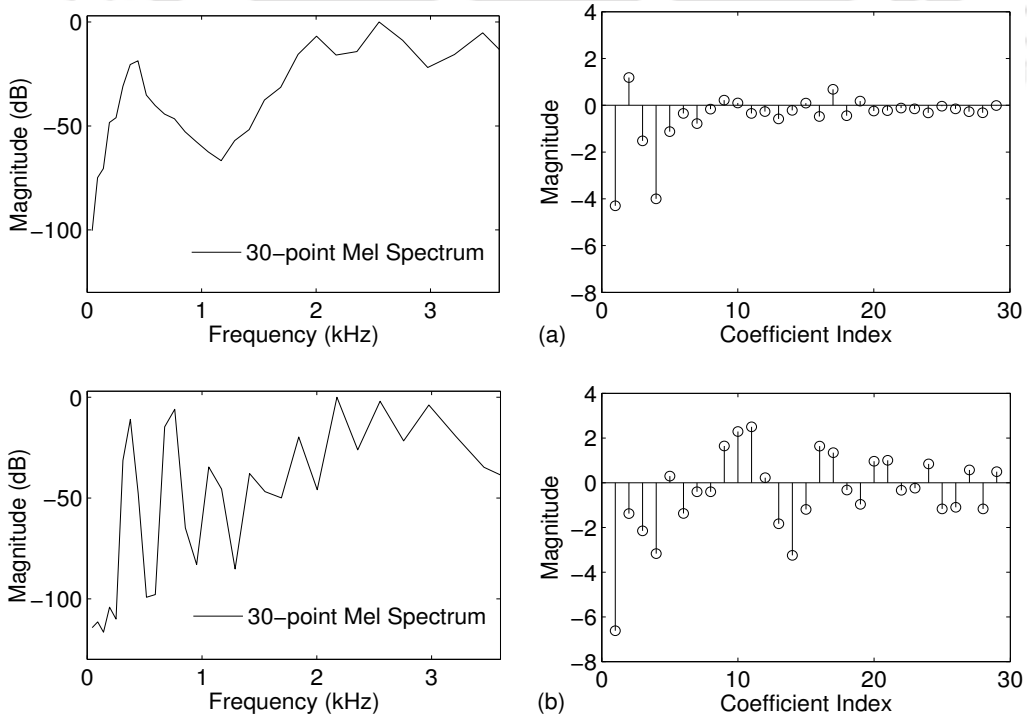


Figure 4.3: Plots of the 30-point Mel spectra (left panel) and their corresponding cepstra (right panel) for central steady-state portions of vowel /IY/ having pitch values of around (a) 100 Hz (b) 300 Hz.

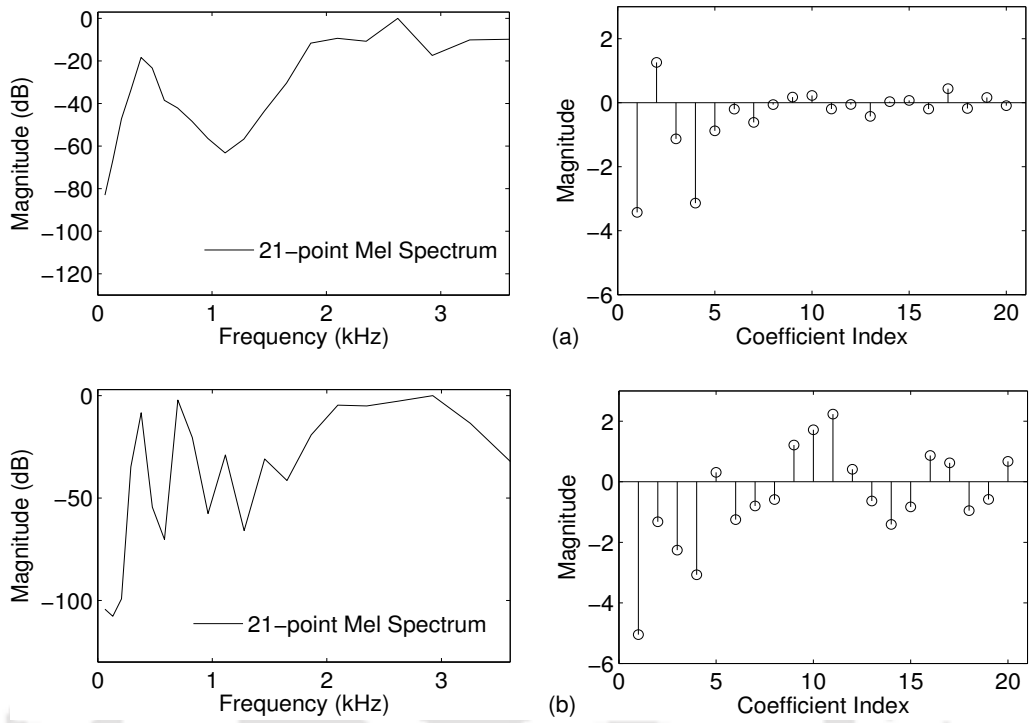


Figure 4.4: Plots of the 21-point Mel spectra (left panel) and their corresponding cepstra (right panel) for vowel /IY/ having pitch values of around (a) 100 Hz (b) 300 Hz.

from the 21-channel Mel filterbank are used for parameterizing a speech frame of 8 kHz sampling rate. Truncation of cepstral features from 21 to 13 dimensions would result in additional smoothing of the 21-point Mel spectrum. Therefore, we further verify that whether the pitch-dependent distortions observed in the Mel spectral envelope of the high pitch signal appear in the smoothed Mel spectral envelope corresponding to the truncated 13-D MFCC features.

The smoothed Mel spectrum corresponding to the truncated 13-D Mel cepstrum is obtained by computing an inverse discrete cosine transform (IDCT) of the 13-D Mel cepstrum ($C_0 - C_{12}$) after appending zeros to the cepstrum. For better exposition of the details in the smoothed Mel spectrum, 115 zeros are appended to the 13-D Mel cepstrum to obtain a 128-point smoothed Mel spectrum (referred to as 'Smoothed'). Figure 4.5 shows the plots of the signals and the 128-point smoothed Mel spectra along with their corresponding linear DFT spectra for central steady-state portions of vowel /IY/ having pitch values of around 100 Hz, 220 Hz and 300 Hz. It is to note that significant pitch-dependent distortions appear in the smoothed Mel spectral envelope particularly at the lower frequencies (below 1 kHz) for the 300 Hz pitch signal which are similar to those observed in case of the 21-point Mel spectrum obtained as the output of the Mel filterbank for the high pitch signal in

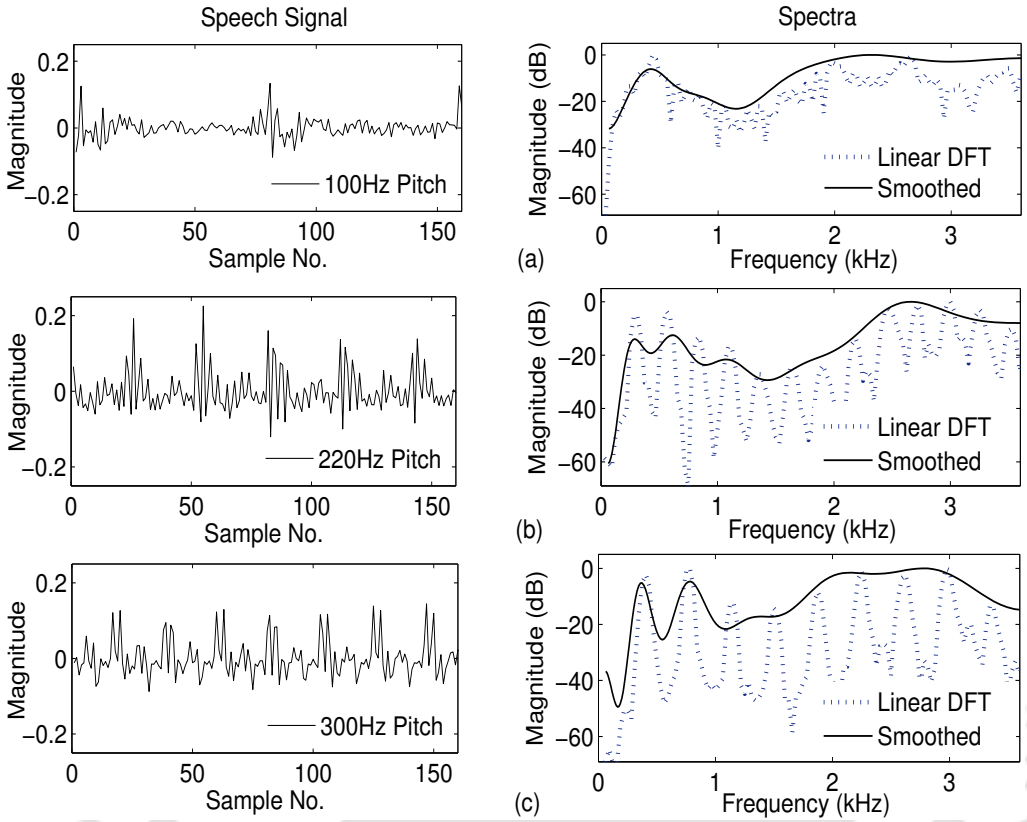


Figure 4.5: Plots of the signals and the 128-point smoothed Mel spectra (referred to as 'Smoothed') along with their corresponding linear DFT spectra for vowel /IY/ having pitch values of around (a) 100 Hz (b) 220 Hz (c) 300 Hz.

Figure 4.4. On the other hand, no such pitch-dependent distortions are noted in the smoothed Mel spectral envelope of the 100 Hz signal which rather appears to be sufficiently smoothed out similar to the 21-point Mel spectrum of the low pitch signal noted in Figure 4.4. Further, on comparing the 128-point smoothed Mel spectra corresponding to signals having pitch values of around 100 Hz, 220 Hz and 300 Hz, it is also noted that the extent of the pitch-dependent distortions along the frequency range and their magnitude are increasing with increasing pitch of the signals.

In a Mel filterbank, the bandwidths of the filters are chosen to approximate the critical bandwidth phenomena observed in the psychoacoustic studies for human auditory perception [2]. The comparison of the center frequencies and the critical bandwidths for human auditory perception as proposed by Zwicker [2] with those of the 21-channel Mel filterbank as per HTK implementation for 4 kHz signal bandwidth are given in Table 4.1. The non-uniform filters in the Mel filterbank are intended to smooth out the pitch harmonics in the linear DFT spectrum to capture the spectral envelope characterizing the vocal filter in the resulting Mel spectrum. In case of low pitch signals, the Mel filterbank appears

Table 4.1: The center frequencies and the critical bandwidths for human auditory perception as proposed by Zwicker [2] and the center frequencies along with the corresponding bandwidths of all filters of a 21-channel Mel filterbank as per the HTK implementation for 4 kHz signal bandwidth.

Filter No.	Critical Bandwidths Proposed by Zwicker		21-channel Mel filterbank as per HTK Implementation	
	Center Frequency	Bandwidth	Center Frequency	Bandwidth
1	50	**100**	63.3	**132**
2	150	100	132.3	144
3	250	100	207.6	157
4	350	100	289.6	172
5	450	110	379.1	187
6	570	120	476.6	204
7	700	140	583.0	222
8	840	150	699.0	242
9	1000	160	825.5	264
10	1170	190	963.4	288
11	1370	210	1113.8	314
12	1600	240	1277.8	343
13	1850	280	1456.7	374
14	2150	320	1651.6	408
15	2500	380	1864.3	444
16	2900	450	2096.1	485
17	3400	550	2348.9	528
18	4000	**700**	2624.6	576
19	4080	900	2925.1	628
20	5800	1100	3252.9	685
21	7000	1300	3610.3	**747**
22	8500	1800	-	-
23	10500	2500	-	-
24	13500	3500	-	-

to effectively smooth out the pitch harmonics as the bandwidth of the narrowest filter in the Mel filterbank is also comparable to their pitch harmonic frequency. This can be further understood by noting the center frequencies and the bandwidths of the filters of the 21-channel Mel filterbank for 4 kHz signal bandwidth given in Table 4.1. As a result, the Mel spectrum and the Mel cepstrum of

the low pitch signal are also noted to contain no significant pitch harmonicity. However, in case of high pitch signal, due to greater separation between the pitch harmonics the smoothing by the typical non-uniform Mel filterbank is not as effective due to the bandwidths of some filters being lesser than the pitch of the signal. This is attributed to cause the undesired pitch-dependent distortions in the Mel spectral envelope in case of high pitch signals. The bandwidths of the filters in the Mel filterbank increase with increasing center frequencies of the filters. This results in increase in the smoothing of the pitch harmonics along the frequency in the spectrum. As a result, the pitch-dependent distortions appear predominantly at low frequencies in the smoothed Mel spectral envelope of the high pitch signals as shown in Figure 4.5.

The occurrence of the pitch-dependent distortions in the Mel spectral envelope for high pitch signals due to insufficient smoothing of the pitch harmonics by the non-uniform Mel filterbank is also validated using a synthetic example. The Mel spectra at the output of the Mel filterbank are computed for synthetically generated pitch harmonic spectra corresponding to pitch values of 100 Hz, 200 Hz and 300 Hz without the effect of the vocal filter and are shown in Figure 4.6. The synthetic pitch harmonic spectra are created by taking linear DFT of impulse trains of different pitch periods after windowing them using Hanning window of 200 points (corresponding to 25 msec speech frame). It is to note that pitch-dependent distortions appear in the Mel spectral envelope for pitch harmonic spectra of higher frequency which increase with increase in the frequency of the pitch harmonic spectra. This verifies that the pitch-dependent distortions appear in the Mel spectral envelope for high pitch signals only due to the insufficient smoothing of the pitch harmonics by the Mel filterbank.

Despite the appearance of some pitch-dependent distortions in the low frequency region in the Mel spectral envelope of high pitch real signal, the corresponding cepstrum does not appear to contain any pitch-dependent harmonicity. This is attributed to the constant-Q type filters in the Mel filterbank. The bandwidths of the Mel filters increase with increasing center frequencies of the filters. As a result, the regularity of the pitch harmonics in the resulting spectrum is disturbed across the entire frequency range. So, in the Mel filtered cepstra the pitch-related information does not get separated from the effects of the vocal filter. Following these, we argue that the pitch-dependent distortions occurring in the Mel spectral envelope would not show any pitch-dependent harmonicity in the Mel cepstrum but would affect all cepstral coefficients. Therefore, the pitch-related information, if captured, can not be extracted out completely from the Mel cepstrum by cepstral truncation.

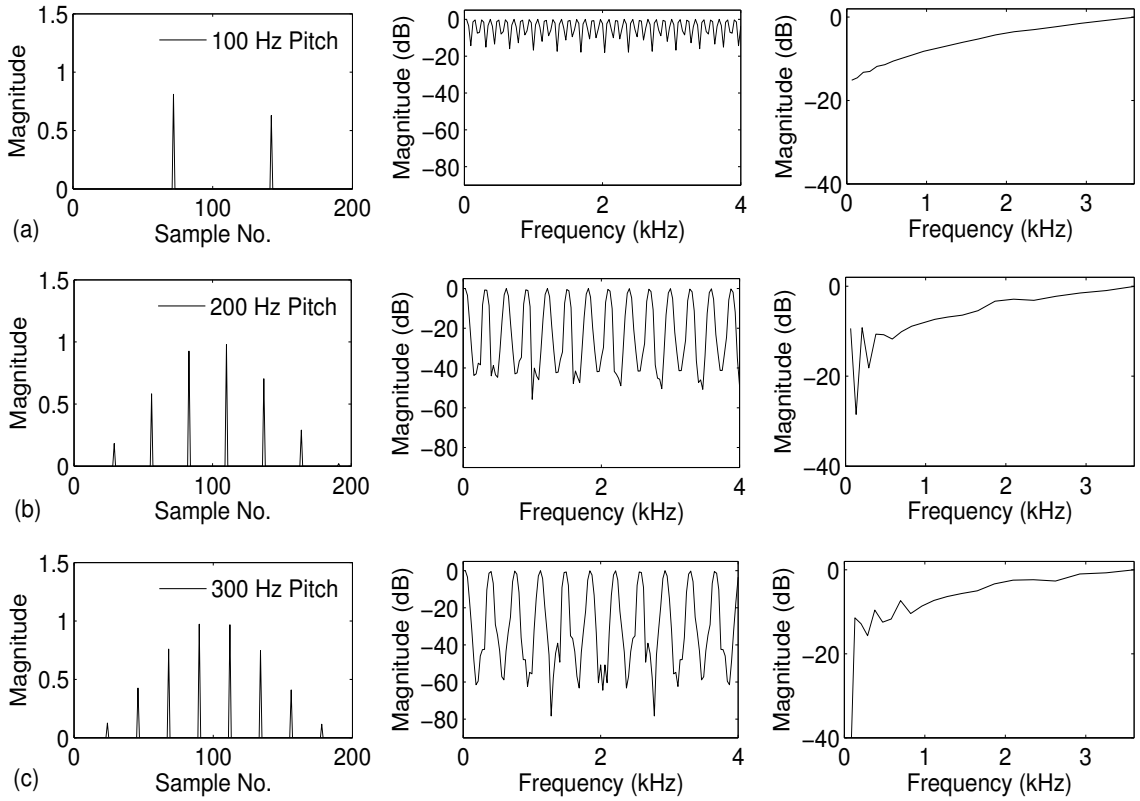


Figure 4.6: Plots showing 21-point Mel spectra (right panel) of the synthetically generated pitch harmonic spectra (middle panel) corresponding to different pitch frequencies (a) 100 Hz (b) 200 Hz (c) 300 Hz. The synthetic pitch harmonic spectra are created by taking linear DFT of impulse trains shown in corresponding left panel. Note that the slope in the Mel spectra is on account of the outputs of the Mel filters not being normalized by their corresponding areas.

4.3 Effect of Pitch-dependent Distortions on MFCCs

In this section, we explore in detail the nature of the effect of the pitch-dependent distortions appearing in the 128-point smoothed Mel spectral envelope for high pitch signal noted in the previous section on the 13-D truncated Mel cepstrum i.e., MFCC ($C_0 - C_{12}$). The 128-point smoothed Mel spectra are computed from their corresponding 13-D truncated Mel cepstra for the central steady-state portions of vowels extracted from low and high pitch signals as described in the previous section. Figure 4.7 shows the plots of the 128-point smoothed Mel spectra for vowels /AE/ and /IY/ having pitch values of around 100 Hz ('low') and 300 Hz ('high') and their corresponding 13-D truncated MFCC.

The effect of the pitch-dependent distortions appearing in the smoothed Mel spectral envelope for high pitch signals on their corresponding cepstra is noted by comparing the MFCCs of the high pitch

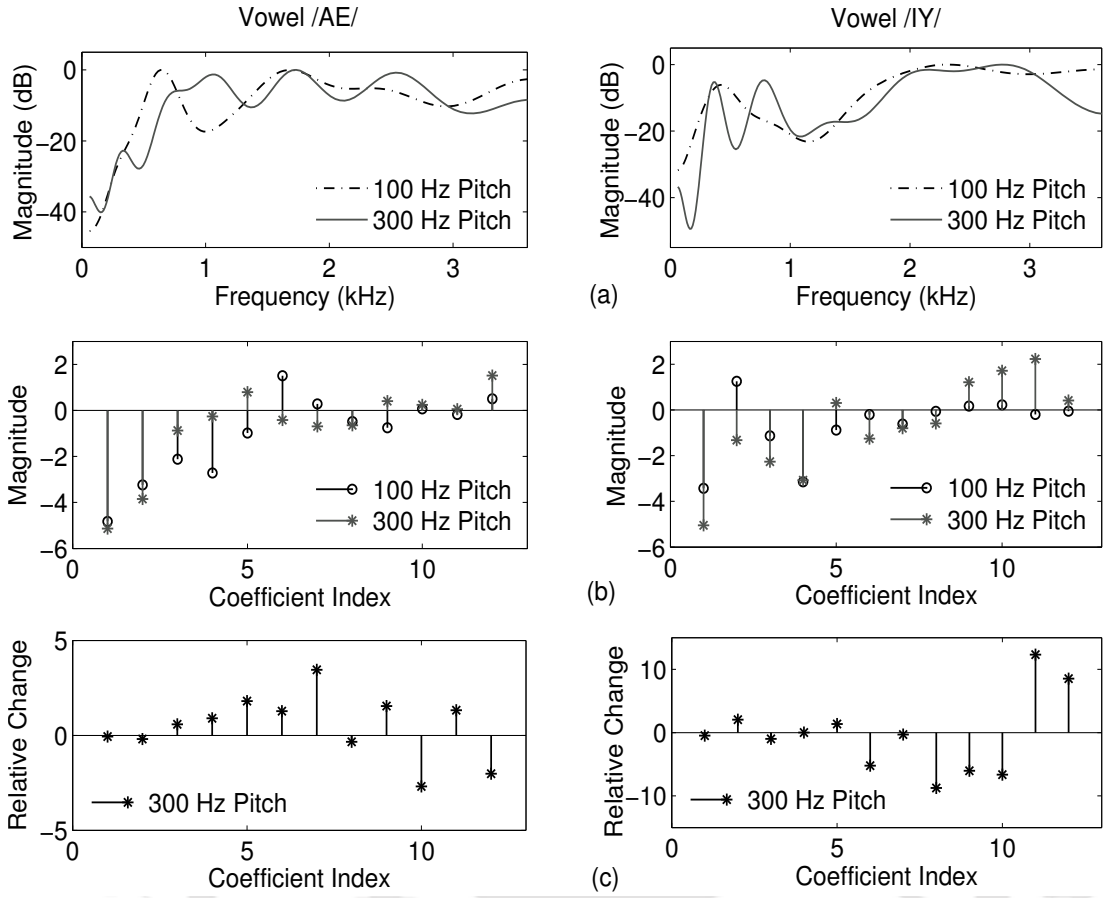


Figure 4.7: Plots for vowels /AE/ and /IY/ having pitch values of around 100 Hz and 300 Hz (a) Smoothed Mel spectra (b) 13-dimensional truncated MFCCs excluding C_0 (c) relative change in each MFCC for the 300 Hz pitch signal with respect to those for the 100 Hz pitch signal.

signals with those of the low pitch signals for same vowels. The relative changes in MFCCs for the high pitch signals with respect to those for the low pitch signals are shown in the bottom panel in Figure 4.7. It is noted that the relative change in MFCCs ($C_1 - C_{12}$) of high pitch signals is more for higher order MFCCs in comparison to that in case of the lower order MFCCs in case of both vowels. From physiological point of view, the higher pitch speech signals correspond to shorter vocal tract lengths. So, the smoothed Mel spectra for the low and high pitch signals may appear to contain some differences due to differences in the vocal tracts in the two cases. Therefore, the changes in the MFCCs of high pitch signals with respect to those of low pitch signals can not be attributed only to the argued pitch-dependent distortions appearing in the smoothed Mel spectral envelope in this case.

Thus, to study the effect of only the pitch-dependent distortions on MFCCs ($C_0 - C_{12}$), the 21-dimensional MFCC features are computed by taking DCT of the 21-point Mel spectra of the synthetic

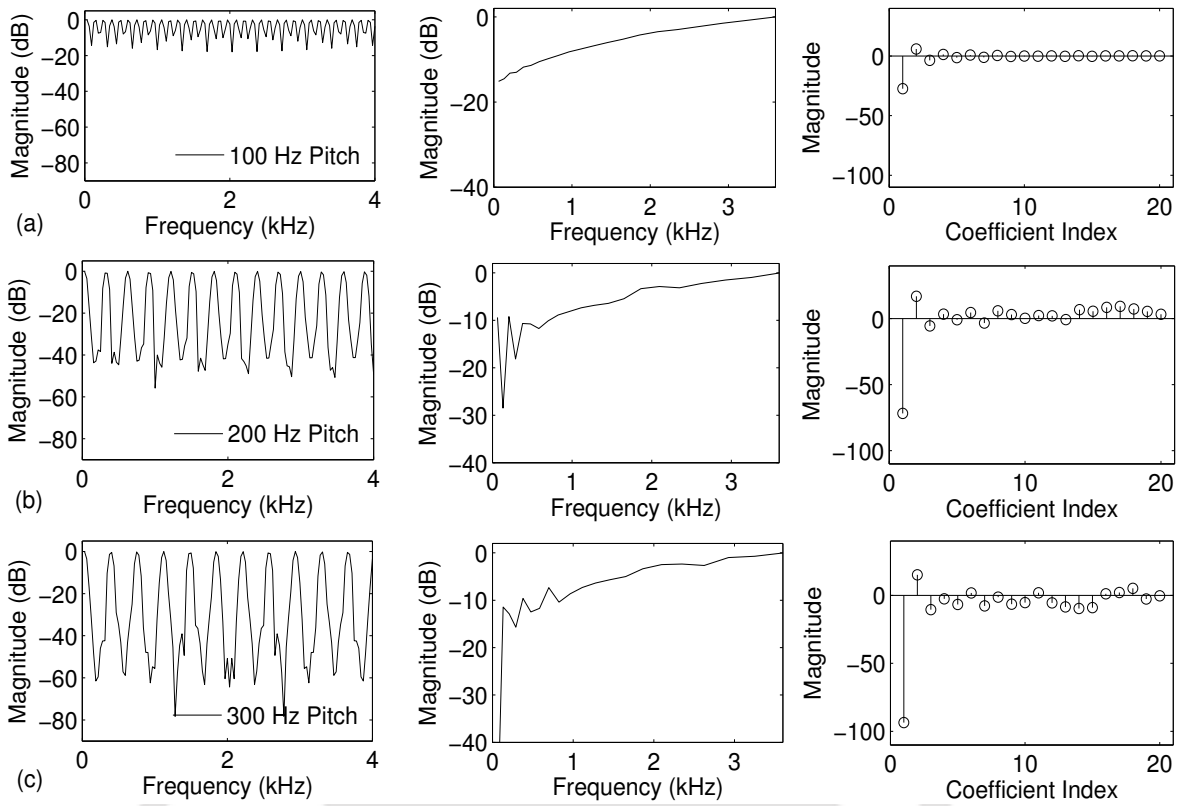


Figure 4.8: Plots of the 128-point linear DFT spectra (left panel), 21-point Mel spectra (middle panel) and their corresponding MFCCs excluding C_0 (right panel) for the synthetically generated pitch harmonic spectra having pitch frequency of around (a) 100 Hz (b) 200 Hz (c) 300 Hz.

pitch harmonic spectra corresponding to different pitch frequencies shown in Figure 4.6. The 21 MFCCs of the synthetic pitch harmonic spectra corresponding to pitch frequencies of around 100 Hz, 200 Hz and 300 Hz along with their corresponding linear DFT spectra and their Mel spectra are shown in Figure 4.8. It is noted that as the pitch frequency is increasing, the pitch-dependent distortions are increasing in the Mel spectra and correspondingly the dynamic range of all coefficients is also increasing in their Mel cepstra.

Further, the relative change in each of the coefficients of MFCC features of the synthetic pitch harmonic spectra corresponding to pitch frequency of 200 Hz and 300 Hz with respect to those of the pitch harmonic spectra corresponding to pitch frequency of 100 Hz are shown in Figure 4.9. It is noted that though due to variation in the pitch frequency all MFCCs are being affected, the relative change is observed more in the higher order coefficients of MFCC features than the lower order MFCCs. As the pitch frequency is increasing the dynamic range of the higher order MFCCs is increasing comparatively more than that for the lower order MFCCs. It is interesting to note that these observations are also

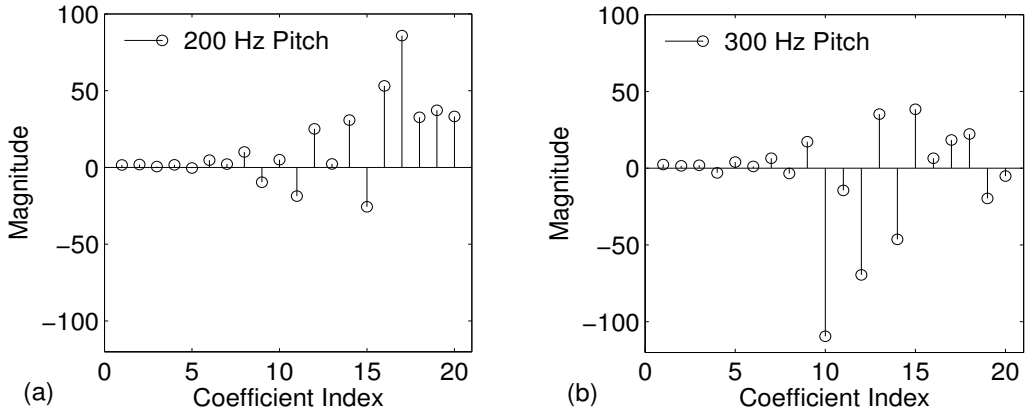


Figure 4.9: Plots showing the relative change in each MFCC ($C_1 - C_{20}$) for the synthetically generated pitch harmonic spectra of different pitch frequencies with respect to those for the synthetic pitch harmonic spectrum having pitch frequency of around 100 Hz (a) 200 Hz (b) 300 Hz.

made in context of the 13-D MFCC features ($C_0 - C_{12}$) which have become *de facto* standard features for all ASR systems. This observation is also consistent with the earlier observed greater relative change in the higher order coefficients than the lower order coefficients of 13-D truncated MFCCs ($C_0 - C_{12}$) of vowel /IY/ having pitch value of around 300 Hz in comparison to MFCCs for vowel /IY/ having pitch value of around 100 Hz extracted from real signals.

Thus, the pitch-dependent distortions in the Mel spectral envelope of high pitch real signals affect all MFCCs but relatively increase the magnitude of the higher order coefficients of 13-D MFCC features to a larger extent in comparison to the lower order coefficients as noted in Figure 4.7. This is attributed for the increase in the variances of the higher order coefficients of 13-D MFCC ($C_0 - C_{12}$) with increase in the pitch of the signals as noted in Section 3.2.1.

It is already known that children's speech have significantly higher pitch values in comparison to those of adults' speech. Children of age from 6-11 years have been reported to have pitch values in the range from 250-350 Hz [27]. Thus, the observations made in the above studies for the 300 Hz pitch signals (both real and synthetic) would be valid for children's speech.

4.4 Summary

In this chapter, the effect of pitch on MFCC features is explored in detail to understand the cause of increase in variances of higher order MFCCs with increase in the pitch of the signals as noted in Section 3.2.1. In summary,

- In ASR, cepstral truncation is employed for estimating the smoothed spectral envelope to derive features for a speech signal. The cepstral truncation is able to remove pitch from the cepstrum corresponding to a uniform filterbank based spectrum only.
- Due to insufficient smoothing of the pitch harmonics by the non-uniform Mel filterbank some pitch-dependent distortions appear in the Mel spectral envelope for the high pitch signals.
- The pitch-dependent distortions observed in Mel spectral envelope for high pitch signals do not show any corresponding pitch-dependent harmonicity in the Mel cepstrum.
- The pitch-related information can not be extracted out completely from the Mel cepstrum by cepstral truncation.
- The pitch-dependent distortions affect all MFCCs ($C_0 - C_{12}$) but relatively cause increase in the magnitude of the higher order coefficients to a larger extent in comparison to the lower order coefficients. As a result, the dynamic range and in turn the variances of the higher order coefficients of MFCC ($C_0 - C_{12}$) features increase with increase in the pitch of the signals as noted in Section 3.2.1.

Motivated by the observed significant effect of pitch on the higher order coefficients of MFCC features in case of high pitch signals, the pitch-robustness of other salient features used in ASR viz., PLPCC and PMVDR is studied in the next chapter to explore their efficacy for children's speech recognition on adults' speech trained models.

5

Pitch-Robustness of Salient ASR Features for Children's ASR

Contents

5.1 Introduction 78
5.2 Efficacy of PLPCC Features for Children's ASR 79
5.3 Children's ASR using PMVDR Features 86
5.4 Summary 90

5.1 Introduction

The most commonly used features for ASR are the MFCC features. Irrespective of the acoustic characteristics of a speech signal, MFCC features are used to parameterize both adults' and children's speech signals. However, the standard approach for MFCC feature extraction has in general been optimized for adults' speech. Since there are large differences in various acoustic correlates of speech for adults' and children's speech, the default MFCC features used for adults' ASR might not be suitable for parameterizing children's speech.

One of the earliest studies which explored different features for children's ASR is reported in [48]. In that study, the children's ASR performances were computed using MFCC features and LPCC features of different model orders. Though it was noted that children's ASR performance improves using LPCC features with lower model order, even greater improvement in the performance was observed using MFCC features. Since then, predominantly the use of only MFCC features has been reported in literature for children's ASR on adults' speech trained models as well. However, in Chapter 3, it is noted that MFCC features are significantly affected by the pitch variations across speech signals, especially in case of high pitch signals. This behavior is also experimentally verified later in Chapter 4. As a result, significant degradation could occur in the children's ASR performance on adults' speech trained models using MFCC features. In recent literature, the use of PLPCC and PMVDR features has been reported for recognizing children's speech on children's speech trained models [11, 63].

The PLPCC features are obtained through cepstral analysis of the perceptual linear prediction (PLP) [122] coefficients. In PLP analysis of speech, a perceptually motivated filterbank similar to the one applied in MFCC feature extraction is applied on the speech spectrum to smooth out the pitch harmonics. Further, the cubic-root amplitude compression is done on the pre-emphasized critical band spectrum to reduce the spectral amplitude variation. In [122], the PLP based spectral analysis has been shown to give better modeling for children's speech than LP based spectral analysis. Many studies have reported either comparable or slightly better performance for adults' ASR with PLPCC features than that obtained with MFCC features under matched but clean condition while significantly better under noisy condition [123–126]. The improvements in the ASR performances with PLPCC features in comparison to MFCC features are mainly attributed to the reduction in the spectral amplitude variation of the critical-band spectrum due to equal-loudness pre-emphasis and cubic-root amplitude compression [122].

On the other hand, in [127], the PMVDR features have been proposed based on the minimum variance distortionless response (MVDR) [128] spectral analysis of speech. The PMVDR features incorporate perceptual warping along with the MVDR spectral estimator. MVDR obtains the power spectrum estimates by using data-dependent bandpass filters [129]. This helps the MVDR spectrum to accurately model the peaks in the speech spectrum by successfully connecting the spectral peaks to form the spectral envelope [128]. The MVDR spectrum is shown to provide better spectral modeling compared to the LP spectrum especially for medium and high pitched speech signals [128]. MFCC features have also been shown to give improved adults' matched ASR performance using the MVDR spectrum in comparison to that obtained using the DFT spectrum [130]. This is attributed to the use of frequency-dependent and data-dependent bandpass filters used in MVDR based spectral analysis unlike the DFT based analysis where fixed bandpass filters are used regardless of the characteristics of the incoming signal. The MVDR spectral analysis based PMVDR features have also been reported to give comparable or slightly better ASR performance than MFCC features (both DFT-based and MVDR-based) for adults' speech under matched but clean condition while significantly better under noisy condition [127].

Motivated by these, in this chapter, we study the pitch-robustness of PLPCC and PMVDR features to explore their efficacy for children's speech recognition on adults' speech trained models in comparison to MFCC features. All speech recognition evaluations are demonstrated on a limited vocabulary connected digit recognition task.

The rest of the chapter is organized as follows: In Section 5.2, the effect of pitch is explored on PLPCC features while the pitch-robustness of PMVDR features is explored in Section 5.3 for children's speech recognition on adults' speech trained models. Finally, the observations made in this chapter are summarized in Section 5.4.

5.2 Efficacy of PLPCC Features for Children's ASR

In this section, first the effect of pitch variations across speech signals is analyzed on PLPCC features using TIMIT data in comparison to that noted in case of MFCC features in Section 3.2.1. Then, to quantify the effect of pitch on PLPCC features in context of ASR, the recognition results are evaluated using TIDIGIT corpus for children's speech on adults' speech trained models. In this study, all PLPCC features are computed using the HTK toolkit [89]. The 13-dimensional ($C_0 - C_{12}$) base

PLPCC features are derived using 12^{th} order LP analysis and a 21-channel triangular Mel filterbank. The LP order chosen for PLPCC feature extraction is consistent with the literature [131]. In addition to the base features, their first and second order temporal derivatives, computed over a span of 5 frames, are also appended making the final feature dimension as 39. The PLPCC feature computation process as proposed by Hermansky in [122] is described in detail in Appendix C. Cepstral mean subtraction is applied to the features.

5.2.1 Effect of Pitch on PLPCC Features

The same set of speech frames from 'low' (100-125 Hz) and 'high' (200-250 Hz) pitch group speech signals from the TIMIT database are selected which are used for analyzing the effect of pitch on MFCC features in Section 3.2.1. The central steady-state portions of 7 different vowels present in the different pitch group signals are extracted and their corresponding PLPCC features are computed. Approximately, 2000 frames are used for each vowel. The plot of mean and the bar-plot showing variance of each of the 12 static coefficients ($C_1 - C_{12}$) of PLPCC features of signals belonging to low and high pitch groups for a representative vowel /IY/ are shown in Figure 5.1(a). For ease of comparison, the plots corresponding to MFCC features for the same are shown in Figure 5.1(b). It is noted that the variance of the higher dimensions of PLPCC features of the high pitch group signals is significantly larger than that of the low pitch group signals as observed in case of MFCC features in Section 3.2.1. However, the relative increase in the variances of the higher order coefficients of the features with increase in the pitch of the signals is lesser in case of PLPCC features than in case of MFCC features. The plot of mean and the bar-plot showing variance of each of the 12 static coefficients ($C_1 - C_{12}$) of PLPCC features for 200-250 Hz pitch group signals before and after their average pitch transformation by factor of 0.7 for vowel /IY/ are also shown in Figure 5.1(a). For ease of comparison, the plots corresponding to MFCC features for the same are also shown in Figure 5.1(b). From these plots, it is noted that on pitch reduction the variances of the higher dimensions of PLPCC features also reduce considerably like in case of MFCC features as noted in Section 3.2.1.

In case of MFCC features, the increase in the variances of the higher order coefficients with increase in the pitch of the speech signals is attributed to the occurrence of the pitch-dependent distortions in the smoothed Mel spectral envelope. Therefore, the effect of pitch variations across speech signals is further explored on the smoothed spectrum corresponding to PLPCC features. The smoothed spectrum corresponding to PLPCC features is obtained by converting the cepstral coefficients to their

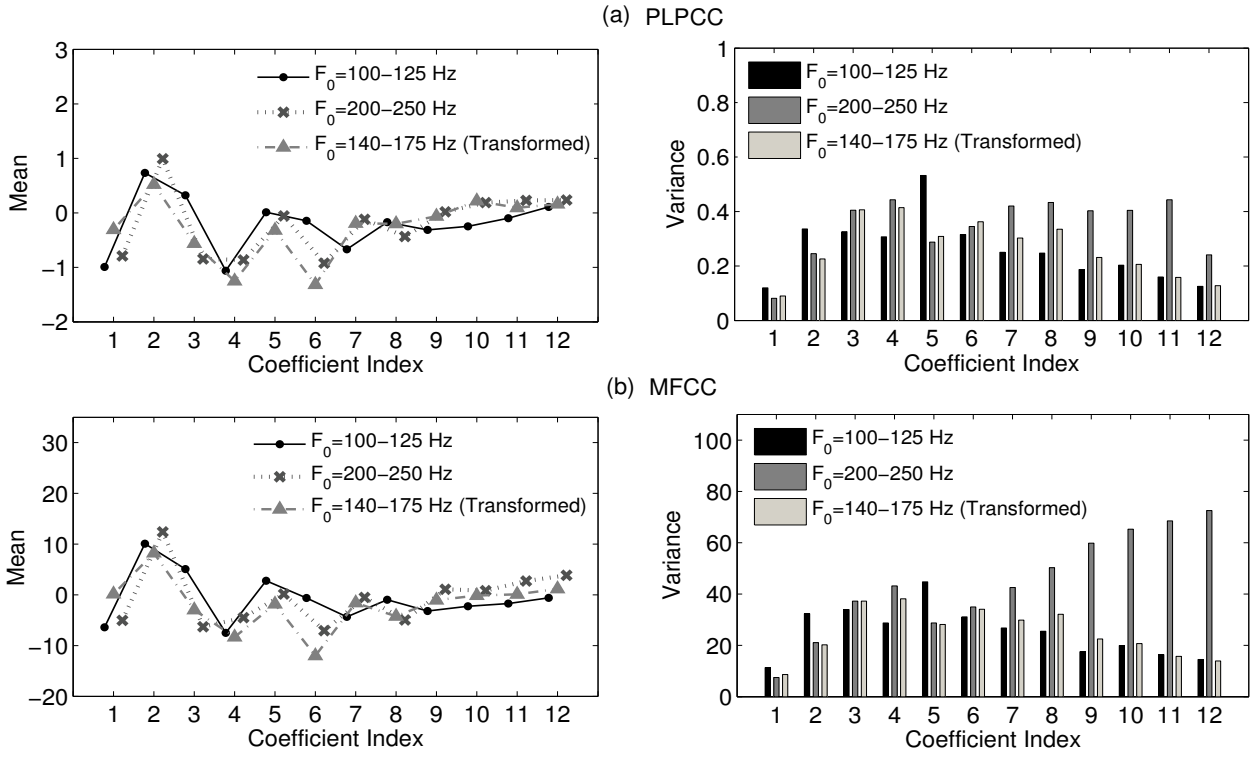


Figure 5.1: Plots showing mean (left panel) and bar-plots showing variance (right panel) of each of the coefficients ($C_1 - C_{12}$) of (a) PLPCC features and (b) MFCC features for signals of different pitch groups: 100-125 Hz, 200-250 Hz and 200-250 Hz transformed to 140-175 Hz for vowel /IY/.

corresponding LP coefficients and then obtaining the magnitude response of the all-pole filter for those LP coefficients. Figure 5.2 shows the smoothed spectra corresponding to PLPCC features along with the linear DFT spectra for central steady-state portions of vowel /IY/ having pitch values of around 100 Hz, 220 Hz and 300 Hz. Similar to the case of MFCC features, some pitch-dependent distortions are noticed at lower frequencies (below 1.5 kHz) in the smoothed spectral envelope corresponding to PLPCC features for 220 Hz and 300 Hz pitch signals when compared with that of the 100 Hz pitch signal. The pitch-dependent distortions in the smoothed spectral envelope are also noted to increase with increase in average pitch of signals. These pitch-dependent distortions in the smoothed spectral envelope corresponding to PLPCC features are attributed to the earlier observed increase in the variances of PLPCC with increasing pitch of the signals. The computation of PLPCC features involves the same filterbank for perceptual warping [122] as used in the computation of MFCC features. So, we hypothesize that the possible cause of the distortions in the smoothed spectral envelope in case of PLPCC features is same as that attributed for MFCC features, i.e., the insufficient smoothing of the pitch harmonics by the filterbank particularly at low frequencies by the lower order filters of the

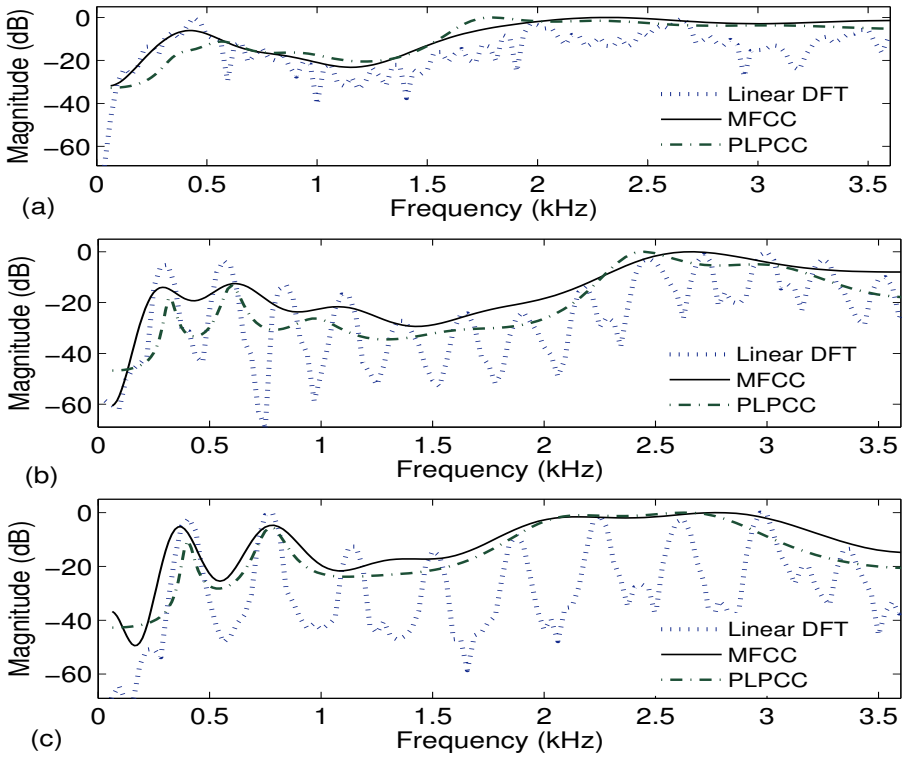


Figure 5.2: Plots of smoothed spectra corresponding to PLPCC features along with the linear DFT spectra for vowel /IY/ having pitch values of around (a) 100 Hz (b) 220 Hz (c) 300 Hz.

filterbank having bandwidth of nearly 100 Hz only. However, on observing the dynamic range of the pitch-dependent distortions in the smoothed spectral envelope, it is to note that the magnitude of the pitch-dependent distortions in the smoothed spectral envelope corresponding to PLPCC features is lesser than those observed in case of MFCC features. This is attributed to the relatively lesser increase in the variances of higher order coefficients of PLPCC features with increase in pitch of the signals than in case of MFCC features.

To quantify the degree of difference in the effect of pitch variations on PLPCC features in comparison to MFCC features, the MD [102] is measured for the PLPCC features of the original low and high pitch group signals and the high pitch group signals with pitch transformed to 140-175 Hz pitch range with respect to the distribution of PLPCC features of the 75-100 Hz pitch group signals. The MD is computed for each PLPCC $(C_1 - C_{12})$ feature vector of all pitch groups using Eqn. 5.1 :

$$\mathtt{MD}(\mathbf{x}, \mu_L) = \sqrt{(\mathbf{x} - \mu_L)^T \mathbf{\Sigma}_L^{-1} (\mathbf{x} - \mu_L)} \tag{5.1}$$

where, $\mathbf{x}$ represents the PLPCC $(C_1 - C_{12})$ feature vector whose distance is to be computed. μ_L

Table 5.1: Mean and variance of the squared Mahalanobis distances (MD) of the PLPCC ($C_1 - C_{12}$) and MFCC ($C_1 - C_{12}$) (taken from Section 3.2.1 for ease of comparison) features of the original signals of 100-125 Hz and 200-250 Hz pitch groups and the transformed signals with pitch transformation from 200-250 Hz to 140-175 Hz pitch range from the distribution of PLPCC features of 75-100 Hz pitch group signals for different vowels.

Vowel	Features	Squared MD (Mean / Variance)		
		Pitch Group		
		100-125 Hz (Original)	200-250 Hz (Original)	140-175 Hz (Transformed from original 200-250 Hz)
/AE/	PLPCC	12.4 / 60.4	50.2 / 711.9	36.1 / 253.2
	MFCC	12.4 / 60.2	65.4 / 2553.1	35.3 / 282.7
/IY/	PLPCC	13.5 / 69.5	43.6 / 509.7	36.7 / 253.5
	MFCC	13.5 / 85.3	59.2 / 1735.4	34.7 / 246.8

denotes the mean and $\mathbf{\Sigma}_L$ denotes the diagonal-covariance of the distribution of PLPCC features of 75-100 Hz pitch group signals. The mean and variance of the squared MD of the different pitch group signals for /AE/ and /IY/ vowels corresponding to PLPCC ($C_1 - C_{12}$) features are given in Table 5.1.

From Table 5.1, it is noted that the mean and variance of squared MD of PLPCC features of the high pitch group signals with respect to the 75-100 Hz pitch group distribution are significantly larger than those in case of the low pitch group signals for both vowels. The larger mean and variance of squared MD for the high pitch group signals indicates poor classification performance of the speech recognition models trained on the 75-100 Hz pitch group signals for those signals which verifies the impair effect of pitch on PLPCC features of high pitch group signals. However, the mean and variance of squared MD of PLPCC features of the high pitch group signals are far lower than those of MFCC features of those signals for both vowels. The ratio of variance of squared MD of PLPCC features of vowel /AE/ from low and high pitch group signals turns out to be 11.9 while for MFCC features it becomes as high as 42.5. Also, after explicit pitch reduction of the high pitch group signals to 140-175 Hz pitch range, the degree of reduction in the mean and variance of squared MD is more in case of MFCC features in comparison to PLPCC features. For /AE/ vowel, the variance of squared MD of PLPCC features of the high pitch group signals reduces by factor 3 after their pitch being reduced by factor of 0.7 while in case of MFCC features it reduces greatly by a factor of 9.

Thus, in comparison to MFCC features, PLPCC features are little more pitch-robust and therefore,

are effected by pitch variations to a comparatively lesser extent. The relatively lesser impact of pitch differences on PLPCC features than MFCC features is hypothesized to be due to the incorporation of the equal-loudness pre-emphasis and the cubic-root amplitude compression in PLPCC feature computation [122]. Motivated by this, we further explore the efficacy of PLPCC features for children's ASR performance in comparison to MFCC features.

5.2.2 Children's Speech Recognition using PLPCC Features

The baseline recognition performance (in WER) of the connected digit recognizer on the adults' test set ADts for PLPCC features is 0.47% which is slightly inferior to the baseline performance of 0.43% obtained with MFCC features for test set ADts. This trend in the adults' matched ASR performances with PLPCC and MFCC features is consistent with that already reported in [132]. For the children's test set CHts1, the baseline performances for PLPCC features as well as that for MFCC features (for ease of comparison) are given in Table 5.2 along with a breakup for different pitch groups. The average pitch (F_o) of the signals is estimated using the ESPS tool available in the Wavesurfer software package [88] as described in Section 2.2. For children's speech recognition on adults' speech trained models, a slightly better performance is obtained with PLPCC features in comparison to that with MFCC features. This trend in the performances with PLPCC and MFCC features is consistent with that reported in [133]. The slightly better children's ASR performance with PLPCC in comparison to MFCC features could be attributed to the incorporation of the equal-loudness pre-emphasis and the

Table 5.2: Performance for children's test set CHts1 (with breakup for different pitch groups) with and without pitch normalization for default PLPCC and MFCC (taken from Section 3.3.1.1 for ease of comparison) features. The 95% confidence interval for the performance for CHts1 data set is ±0.39 [for $F_o < 250$ Hz, $250 \leq F_o < 300$ Hz and $F_o \geq 300$ Hz pitch groups ±0.39, ±0.79 and ±3.37, respectively].

Features	Condition	% WER			
		All F_o Values (7,772)	$F_o <$ 250 Hz (5,224)	250 Hz $\leq F_o <$ 300 Hz (2,346)	$F_o \geq$ 300 Hz (202)
PLPCC	Baseline	10.61	6.70	15.52	33.09
	Norm.	10.05	6.44	14.62	30.61
MFCC	Baseline	11.37	6.54	17.47	39.03
	Norm.	9.64	6.02	14.24	30.11

cubic-root amplitude compression in PLPCC feature computation which help reducing any acoustic mismatch unlike MFCC features.

The children's speech have much higher pitch values compared to those of the adult's speech [27]. Motivated by the reduction in the variances of the features with reduction in the pitch of the signals as observed in Section 5.2.1, the average pitch of the children's speech signals of test set CHts1 is modified towards the pitch range of the adults' data of the training set ADtr using the PSTS [1] method as described in Section 2.4.1.1 for reducing the pitch mismatch. For determining the appropriate transformations of pitch values of each of the children's test speech signals, a ML grid search is done similar to the one commonly used for ML-based speaker normalization [55]. The 8-point grid search for determining the optimal pitch value for the test signals involves the original signal and its seven pitch transformed versions with transformed pitch values ranging from 70-250 Hz in steps of 30 Hz. This range of transformed pitch values is chosen considering the pitch ranges of the adults' training sets and the children's test sets. The speech recognition performances of the children's test set CHts1 after explicit pitch normalization for PLPCC features and MFCC features (for ease of comparison) are also given in Table 5.2 along with a breakup for different pitch groups. To study the effect of the pitch mismatch of the children's test data with respect to the pitch range of the adults' training data, the pitch groups of $F_o < 250$ Hz, 250 Hz $\leq F_o < 300$ Hz and $F_o > 300$ Hz are chosen for the children's test set. The pitch group $F_o < 250$ Hz matches with the pitch range of the adults' training speech data.

From Table 5.2, it is noted that the explicit pitch normalization has resulted in 5% relative improvement in children's ASR performance with PLPCC features. This performance improvement with pitch reduction is attributed to reduction of the pitch-dependent distortions in the smoothed spectral envelope observed in previous section. On observing the pitch group-wise performances given in Table 5.2, it is noted that with pitch normalization consistent improvements are obtained for different pitch groups i.e., higher pitch groups show greater improvements for PLPCC features. However, it is to note that after explicit pitch normalization the recognition performance obtained for children's speech using PLPCC features is slightly inferior with relatively lesser improvements for different pitch groups in comparison to those in case of MFCC. This is attributed to the equal-loudness pre-emphasis involved in PLPCC feature computation to approximate the nonequal sensitivity of human hearing at different frequencies and to simulate the sensitivity of hearing at 40 dB level [122]. The approximation

of the equal-loudness pre-emphasis function is given in Eqn. 5.2.

$$E(\omega) = [(\omega^2 + 56.8 \times 10^6)/\omega^4]/[(\omega^2 + 6.3 \times 10^6)^2 \times (\omega^2 + 0.38 \times 10^9)] \quad (5.2)$$

The pre-emphasis function deemphasizes the 0.4-1.2 kHz spectral region by 6dB compared to the 1.2-3.1 kHz spectral region. Though it significantly reduces the effect of pitch-dependent variations (appearing mainly below 1.5 kHz), this may also lead to de-emphasis of the relevant spectral information which becomes more obvious when the test data is explicitly pitch normalized or devoid of significant pitch differences. This is attributed to the comparatively lesser improvement in children's ASR performance for $F_o < 250$ Hz pitch group signals in comparison to higher pitch group signals after explicit pitch normalization of children's speech with PLPCC features. The above fact is also supported by an earlier observation that with no significant pitch differences between training and test sets, the performance of the adults' test set is found to be slightly inferior for PLPCC features compared to that for MFCC features.

5.3 Children's ASR using PMVDR Features

In this study, the 'default' base PMVDR features are computed using LP analysis of order 20 chosen after extensive search on adults' test set. The chosen LP order of 20 is consistent with the one reported in [127]. A perceptual warp factor of 0.31 is used for computation of all PMVDR features. The 13-dimensional PMVDR features consist of $C_1 - C_{12}$ coefficients augmented with normalized log frame energy. In addition to the base features, their first and second order temporal derivatives, computed over a span of 5 frames, are also appended making the final feature dimension as 39. Cepstral mean subtraction is also applied. All PMVDR features in this work are computed using the SONIC toolkit [134]. The step-by-step procedure for computing PMVDR features as proposed by Yapanel *et.al.* in [127] is described in Appendix D.

The recognition performance (in WER) of the connected digit recognizer on the adults' test set ADts for the default PMVDR features is 0.45% which is comparable to the performance of 0.43% obtained with the default MFCC features in Section 3.3.1. The slight degradation in the performance with PMVDR features in comparison to MFCC features could be due to the unequal distribution of male and female speakers in the adults' training set ADtr and adults' test set ADts. This is hypothesized observing the differences in the performances for male and female speech with PMVDR features

Table 5.3: Performance for children's test set CHts1 (with breakup for different pitch groups) with default PMVDR and MFCC features with and without explicit pitch normalization of children's speech.

Features	Condition	% WER			
		All F_o Values (7,772)	$F_o <$ 250 Hz (5,224)	250 Hz $\leq F_o <$ 300 Hz (2,346)	$F_o \geq$ 300 Hz (202)
PMVDR	Baseline	11.57	7.54	16.46	36.93
	Norm.	11.03	7.41	15.25	35.69
MFCC	Baseline	11.37	6.54	17.47	39.03
	Norm.	9.64	6.02	14.24	30.11

and MFCC features under clean condition as reported in [127]. For children's speech recognition, the baseline performances for the children's test set CHts1 using default PMVDR features and MFCC features (for ease of comparison) are given in Table 5.3 along with a breakup for different pitch groups. On comparing the performances with both default PMVDR features and MFCC features, it is noted that, unexpectedly, the performance with default PMVDR features is slightly inferior to that using MFCC features for overall children's test set CHts1. However, on comparing the pitch group-wise performances it is noted that though for signals with pitch values greater than 250 Hz PMVDR features perform better than MFCC features, for signals belonging to pitch group of less than 250 Hz PMVDR features perform significantly inferior to MFCC features.

In order to assess the pitch-robustness of the default PMVDR features in comparison to that of MFCC features, the children's ASR performances are evaluated after explicit pitch normalization of children's speech. The average pitch of each of the children's speech signals of test set CHts1 is explicitly modified towards the pitch range of the adults' training set ADtr using PSTS method [1] as described in Section 2.4.1.1 for reducing the pitch mismatch. For determining the appropriate transformations of pitch values of each of the children's test speech signals, a ML grid search is done as described in Section 5.2.2. The speech recognition performances of the children's test set CHts1 after explicit pitch normalization for PMVDR features and MFCC features (for ease of comparison) are also given in Table 5.3 along with a breakup for different pitch groups. It is noted that relative improvements of 4.6% and 15% are obtained in ASR performance for children's test set CHts1 after explicit pitch normalization of children's speech with default PMVDR and MFCC features, respectively. However, the children's ASR performance obtained with PMVDR features after explicit pitch

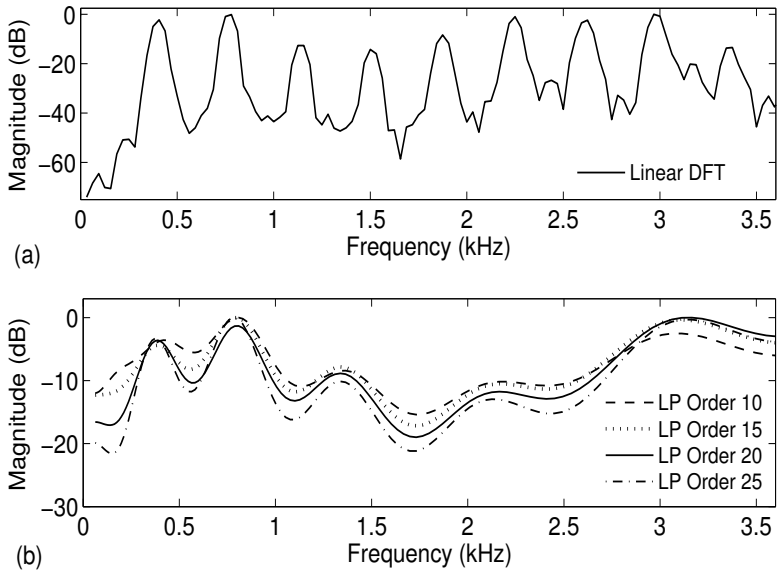


Figure 5.3: Plots for vowel /IY/ having pitch value of around 300 Hz (a) Linear DFT spectrum (b) Smoothed spectra corresponding to PMVDR features computed using various values of LP order.

normalization of children's speech is significantly inferior to that obtained in case of MFCC features across all pitch groups.

Motivated by the inferior performance of the default PMVDR features in comparison to the default MFCC features for children's ASR both with and without explicit pitch normalization, we further explore PMVDR features for children's speech recognition. It has already been shown in [128] that a speech signal having lesser number of pitch harmonics requires lower LP model order for MVDR-based spectral envelope estimation. In order to understand the effect of LP model order on the spectra corresponding to PMVDR features for high pitch signals, the smoothed spectra corresponding to PMVDR features of different LP orders for vowel /IY/ having pitch value of around 300 Hz are shown in Figure 5.3. The smoothed spectrum corresponding to PMVDR features is derived by computing a 128-point IDCT of the 13 base feature coefficients. It is noted that as the LP order decreases the spectral envelope corresponding to PMVDR features becomes smoother. Since children's speech have much higher pitch values than those of adults' speech, it is worth exploring the efficacy of PMVDR features computed using LP orders lower than those used in case of adults' speech for children's speech. The speech recognition performances of the baseline adults' speech trained models are evaluated for children's test set CHts1 using PMVDR features corresponding to different LP orders varying from 10 to 25 in steps of 5 for children's speech and are given in Table 5.4.

Table 5.4: Performances for children's test set CHts1 (with breakup for different pitch groups) with PMVDR features corresponding to various values of LP orders.

LP Order	% WER			
	All F_o Values (7,772)	$F_o <$ 250 Hz (5,224)	250 Hz $\leq F_o <$ 300 Hz (2,346)	$F_o \geq$ 300 Hz (202)
10	9.52	6.60	13.18	26.52
15	9.34	5.92	13.59	29.86
20	11.57	7.54	16.46	36.93
25	13.90	9.35	19.59	40.52

From Table 5.4, it is observed that the best recognition performance for overall children's test set CHts1 with PMVDR features is obtained for LP order of 15 which is 19% relative improvement over that of the default PMVDR features. Further on looking the pitch group-wise performances, it is noted that this significant improvement with reduction in the LP order is a result of the improvement across all pitch ranges. Also, it is worth noting that the best performances for below and above 250 Hz pitch group signals correspond to a lower LP order of 15 and 10, respectively. Therefore, for children's speech PMVDR features should be computed with LP order lower than that used for adults' speech. This improves the performance of children's ASR performance using PMVDR features by reducing the spectral mismatch between adults' and children's speech spectra due to differences in their pitch harmonics. Taking into fact that both adults' training and test sets have unequal number of utterances from male and female speakers, the slight degradation in the baseline ASR performance for adults' test set using PMVDR features in comparison to MFCC features could also be attributed to the slightly improper choice of LP order of 20 for PMVDR feature computation. Henceforth, the features computed using LP order of 15 are referred to as the 'optimized' PMVDR features.

Further, to assess the pitch-robustness of the optimized PMVDR features, children's ASR performance is evaluated on adults' speech trained models after explicit pitch normalization of children's speech and are given in Table 5.5 along with the pitch group-wise breakup. It is noted that no significant improvement is obtained in the children's ASR performance after explicit pitch normalization of children's speech with optimized PMVDR features for any pitch group signals. This shows that with suitable optimization of model order PMVDR features are more pitch-robust in comparison to default

Table 5.5: Performances for children's test set CHts1 (with breakup for different pitch groups) using optimized PMVDR features and default MFCC features with and without explicit pitch normalization of children's test speech.

Features	Condition	% WER			
		All F_o Values (7,772)	$F_o <$ 250 Hz (5,224)	250 Hz $\leq F_o <$ 300 Hz (2,346)	$F_o \geq$ 300 Hz (202)
Optimized PMVDR	w/o Norm.	9.34	5.92	13.59	29.86
	w/ Norm.	9.23	5.92	13.27	29.62
MFCC	w/ Norm.	9.64	6.02	14.24	30.11

MFCC features. However, on comparing the children's ASR performance obtained using the optimized PMVDR features for children's speech and that obtained using default MFCC features after explicit pitch normalization of children's speech given in Table 5.5, it is noted that comparable performances are obtained in both cases across all pitch groups. The slight differences in their performances could be attributed to the difference in the smoothing effected by the implicit (constant bandwidth) and explicit (constant-Q) filterbank in PMVDR features and MFCC features, respectively. It is also worth noting here that, the number of operations in computation of PMVDR features are *twice* that for MFCC feature computation (see Table 5 in [127]).

5.4 Summary

In this chapter, the effect of pitch variations is analyzed on other salient features used in ASR viz., PLPCC and PMVDR features to explore their efficacy for children's ASR on adults' speech trained models in comparison to MFCC features. The salient observations made from this study are:

- Similar effect of pitch variations is observed on PLPCC features as that noted in case of MFCC features but to a comparatively lesser extent. As a result, PLPCC features give slightly better baseline ASR performance for children's speech than MFCC features.
- After explicit pitch normalization of children's speech, PLPCC features are found to give comparatively poor children's ASR performance than MFCC features. This is attributed to the equal-loudness pre-emphasis incorporated in the PLPCC feature computation.

- PMVDR features are found to be more pitch-robust than MFCC features after suitable optimization of model order for PMVDR feature computation.
- The children's ASR performance obtained using optimized PMVDR features for children's speech is found to be comparable to that obtained with MFCC features after explicit normalization of children's speech.
- Similar to PMVDR features, in case of PLPCC features also we expect to get improvement in the children's ASR performance by lowering the LP analysis order. This has already been noted by Hermansky in [122].

It is already reported in literature, PMVDR feature computation requires *twice* the number of operations compared to that for MFCC features. Motivated by these, in the following chapters we explore appropriate modifications for MFCC feature computation for improving children's speech recognition on adults' speech trained models without explicit pitch normalization of children's speech.

6

Pitch Normalization by Filterbank Modification

Contents

6.1 Introduction 94
6.2 Mel Filterbank Modification for Pitch Normalization 95
6.3 Proposed Pitch Normalization Algorithm 105
6.4 Combining Proposed Algorithm with VTLN and CMLLR 107
6.5 Summary 109

6.1 Introduction

In Section 3.3, it is found that besides formant frequencies, pitch is the other major source of acoustic mismatch leading to significant degradation in the children's ASR performance on the adults' speech trained models. It is noted that significant improvement is obtained in children's ASR performance with explicit ML-based pitch normalization of children's speech using MFCC features. However, ML-based pitch normalization of the speech signals by explicitly transforming the pitch of the signals to various values in time domain is a computationally expensive procedure and its performance is also subject to the accuracy of the pitch marks. Therefore, in this chapter, an automatic algorithm is proposed for computing the pitch normalized MFCC features rather than doing ML grid search over various features derived after explicit pitch transformation in signal domain for children's speech.

For speaker recognition, where it is important to capture the pitch-related information, some studies have already reported improvements in the performances by reducing the bandwidths of the filters in the Mel filterbank for MFCC feature extraction [135–137]. In Chapter 4, it has been shown that some pitch-dependent distortions appear in the smoothed Mel spectral envelope corresponding to MFCC features for high pitch signals. These pitch-dependent distortions are attributed to the insufficient smoothing of the pitch harmonics in the speech spectrum by the non-uniform Mel filterbank particularly by the lower order filters of the filterbank. On account of the constant-Q type Mel filterbank used in MFCC feature computation, there in no harmonicity in the pitch-dependent distortions appearing in the Mel spectrum. As a result, no harmonicity appears in the Mel cepstrum corresponding to those pitch-dependent distortions. Thus, the effect of these pitch-dependent distortions appears on all Mel cepstral coefficients and can not be eliminated completely by cepstral truncation. Motivated by that, the pitch normalization method proposed in this chapter aims at improving the spectral smoothing to reduce the pitch-dependent distortions appearing in the smoothed Mel spectral envelope. The proposed algorithm involves utterance-specific modification of the Mel filterbank structure by increasing the bandwidths of the filters of the filterbank during MFCC feature extraction for each children's test speech signal.

The proposed modification in the bandwidths of the filters of the Mel filterbank for pitch normalization can also be supported by observing the spectra for different orders of the LP model chosen in case of PMVDR features as shown earlier in Figure 5.3. In PMVDR feature extraction, the model order governs the length of the filter used for estimation of the MVDR spectrum. The reduction of

model order reduces the length of the corresponding filter and, therefore, its bandwidth increases. As a result, the smoothing in the estimated spectrum is increased, thereby reducing the effect of pitch harmonics in speech spectrum on the resultant PMVDR features. From these insights, we explore modifications in the Mel filterbank involved in the computation of MFCC features to effect more smoothing in the resulting spectrum so as to reduce the effect of high pitch in children's speech signals on their recognition performance.

The rest of this chapter is organized as follows: In Section 6.2, two ways of modification of Mel filterbank for the intended pitch normalization are discussed. Subsequently, in Section 6.3, a filterbank based pitch normalization algorithm is proposed. The combination of the proposed algorithm with VTLN and CMLLR are explored in Section 6.4. Finally, the work presented in this chapter in summarized in Section 6.5.

6.2 Mel Filterbank Modification for Pitch Normalization

In this section, normalization of the pitch differences across speech signals by spectral smoothing as effected by increasing the bandwidths of the filters of the Mel filterbank used for MFCC feature computation is explored. Two different approaches are discussed for modification of bandwidths of the filters in the filterbank viz., implicit modification of bandwidths of all filters by modifying the number of overlapping triangular filters in the filterbank and selective modification of bandwidths of the filters in the filterbank.

6.2.1 Implicit Modification of Filter Bandwidths

In MFCC feature computation, the filterbank consists of triangular filters having non-uniform bandwidths with 50% overlapping. In this work, the features are computed using HTK toolkit [89] which simplifies design of the non-uniform filterbank by setting up a uniform filterbank with chosen number of filters in filterbank in the Mel domain and then maps it back to the linear frequency domain. The effect of reducing the number of filters on the center frequencies and the bandwidths of the filters in the Mel filterbank is shown in Table 6.1. It is noted that as the number of filters in the filterbank are reducing, the spacing between the center frequencies of all filters and henceforth the bandwidths of all filters are increasing. The effect of implicitly increasing the bandwidths of the filters in Mel filterbank by reducing the number of filters in filterbank on the spectral smoothing can be understood by observing the smoothed spectra corresponding to MFCC features computed with different number

Table 6.1: The center frequencies (CF) and the corresponding bandwidths (BW) of all constant-Q filters of filterbank for different number of filters in the Mel filterbank as per the HTK implementation. Note that in order to increase the bandwidth of filters with lower CFs, the BW of filters with higher CFs is also increased and this may result in over-smoothing in higher frequency regions of speech spectrum.

Filter No.	No. of filters in filterbank																			
	21 (Default)		20		19		18		17		16		15		14		13		12	
	CF	BW	CF	BW	CF	BW	CF	BW	CF	BW	CF	BW	CF	BW	CF	BW	CF	BW	CF	BW
1	63.3	**132**	66.4	**139**	69.9	**147**	73.8	**155**	78.1	**165**	83.0	**176**	88.5	**188**	94.8	**202**	102.0	**219**	110.4	**238**
2	132.3	144	139.2	152	146.8	162	155.4	172	164.9	183	175.8	197	188.1	212	202.3	230	218.8	251	238.3	276
3	207.6	157	218.8	167	231.4	178	245.5	190	261.5	204	279.6	220	300.4	239	324.5	261	352.7	287	386.3	319
4	289.6	172	306.1	183	324.5	195	345.2	210	368.7	227	395.7	246	426.8	269	463.1	296	506.1	329	557.7	370
5	379.1	187	401.5	200	426.8	215	455.4	232	488.0	252	525.6	275	569.2	303	620.6	336	681.8	377	756.0	428
6	476.6	204	506.1	219	539.4	236	577.2	256	620.6	280	670.8	308	729.6	341	799.3	382	883.2	432	985.7	496
7	583.0	222	620.6	240	663.2	260	711.8	283	767.9	311	833.3	344	910.3	384	1002.3	433	1113.8	495	1251.7	574
8	699.0	242	745.9	263	799.3	286	860.7	313	931.7	346	1015.0	385	1113.8	433	1232.7	492	1378.1	567	1559.5	664
9	825.5	264	883.2	288	949.1	315	1025.2	346	1113.8	384	1218.3	431	1343.1	487	1494.3	559	1680.9	650	1916.0	769
10	963.4	288	1033.4	315	1113.8	346	1207.0	383	1316.2	427	1445.7	482	1601.3	549	1791.3	634	2027.8	744	2328.7	890
11	1113.8	314	1198.0	345	1295.0	380	1408.1	423	1541.2	475	1700.0	539	1892.2	618	2128.6	720	2425.2	853	2806.4	1031
12	1277.8	**343**	1378.1	**377**	1494.3	**418**	1630.3	**468**	1791.3	**528**	1984.5	**603**	2219.8	**697**	2511.4	**818**	2880.6	**977**	3359.6	**1194**
13	1456.7	374	1575.4	413	1713.5	460	1876.0	517	2069.3	587	2302.7	674	2588.8	785	2946.1	928	3402.3	1119	-	-
14	1651.6	408	1791.3	452	1954.6	506	2147.5	572	2378.4	653	2658.6	754	3004.4	884	3439.7	1054	-	-	-	-
15	1864.3	444	2027.8	495	2219.8	557	2447.7	632	2721.9	725	3056.7	843	3472.6	996	-	-	-	-	-	-
16	2096.1	485	2286.7	542	2511.4	612	2779.5	699	3103.7	806	3501.9	943	-	-	-	-	-	-	-	-
17	2348.9	528	2570.2	594	2832.2	674	3146.3	772	3528.2	896	-	-	-	-	-	-	-	-	-	-
18	2624.6	576	2880.6	650	3185.1	741	3551.8	854	-	-	-	-	-	-	-	-	-	-	-	-
19	2925.1	628	3220.5	712	3573.1	815	-	-	-	-	-	-	-	-	-	-	-	-	-	-
20	3252.9	685	3592.6	780	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
21	3610.3	747	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

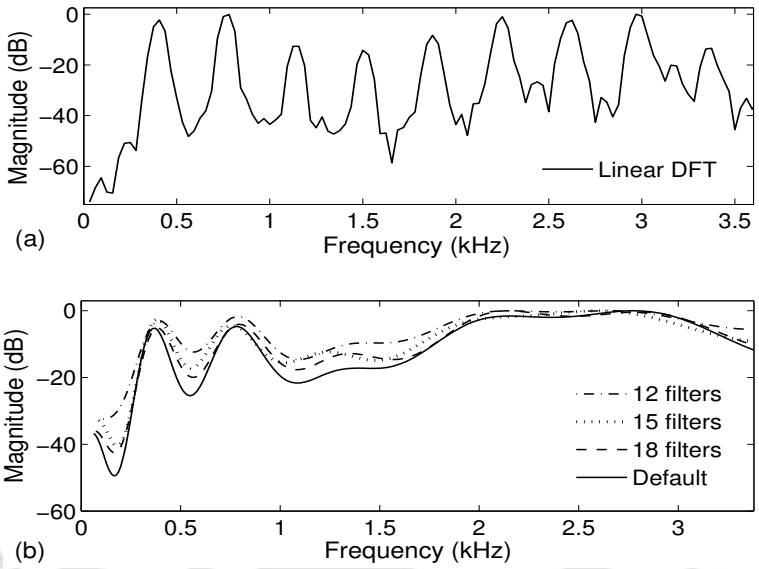


Figure 6.1: Plots for vowel /IY/ having pitch value of around 300 Hz (a) Linear DFT spectrum (b) Smoothed Mel spectra computed using various number of filters in the Mel filterbank.

of filters in the filterbank for vowel /IY/ having pitch value of around 300 Hz as shown in Figure 6.1. It is noted that greater smoothing is obtained as the number of channels (filters) in the filterbank are reduced compared to that in case of the default 21-channel filterbank.

Motivated by that, in this section, we explore the effect of spectral smoothing as effected by implicitly increasing the bandwidths of all filters by reducing the number of filters in the filterbank on the children's ASR performance. The recognition experiments are performed using modified MFCC features due to use of different number of channels in the filterbank for children's speech on the acoustic models trained with default MFCC features for adults' speech. The ten different values chosen for the number of filters in the filterbank for MFCC feature computation of children's speech range from 12 to 21 in steps of 1. The number of filters are reduced only up to 12 in order to keep the number of coefficients in base MFCC feature vectors same for both adults' training data and children's test data. The speech recognition performances of the children's test sets on both the connected digit recognition task and the continuous speech recognition task corresponding to various number of filters in the Mel filterbank for MFCC feature computation along with their breakup for different pitch groups are given in Table 6.2. The average pitch (F_o) of the signals is estimated using the ESPS tool available in the Wavesurfer software package [88] as described in Section 2.2. To study the effect of the pitch mismatch of the children's test data with respect to the pitch range of the adults' training data, the pitch groups of $F_o < 250$ Hz, 250 Hz $\leq F_o < 300$ Hz and $F_o > 300$ Hz are chosen for the children's test set. The

Table 6.2: Performances for children's test sets CHts1 (on connected digit recognition task) and PFts (on continuous speech recognition task) with MFCC features computed using various number of filters in the Mel filterbank along with their pitch group-wise breakup. The 95% confidence interval for the performance for CHts1 data set is ±0.39 [for $F_o < 250$ Hz, 250 Hz $\leq F_o < 300$ Hz and $F_o > 300$ Hz pitch groups ±0.39, ±0.79 and ±3.37, respectively] and for PFts data set is ±1.37 [for $F_o < 250$ Hz, 250 Hz $\leq F_o < 300$ Hz and $F_o > 300$ Hz pitch groups ±1.80, ±1.83 and ±1.61, respectively].

No. of filters in filterbank	% WER							
	CHts1				PFts			
	All F_o Values (7,772)	$F_o <$ 250 Hz (5,224)	250 Hz $\leq F_o <$ 300 Hz (2,346)	$F_o \geq$ 300 Hz (202)	All F_o Values (129)	$F_o <$ 250 Hz (63)	250 Hz $\leq F_o <$ 300 Hz (54)	$F_o \geq$ 300 Hz (12)
Default (21)	11.37	6.54	17.47	39.03	**56.34**	**33.05**	77.25	102.69
20	11.06	6.20	17.31	37.55	56.78	33.89	76.46	106.60
19	10.50	**6.05**	16.30	34.20	56.74	35.75	74.88	101.96
18	10.14	6.13	15.24	32.71	57.02	35.87	**74.43**	106.85
17	10.08	6.10	14.98	33.95	57.94	37.65	76.01	99.02
16	**10.05**	6.52	14.39	31.47	59.58	38.87	77.94	101.96
15	10.09	6.61	14.45	**30.48**	61.83	44.87	76.60	97.80
14	10.20	6.84	**14.28**	31.23	65.96	50.84	79.37	**96.82**
13	11.24	7.96	15.28	31.23	69.67	53.95	84.55	97.07
12	13.34	9.77	17.73	35.07	81.47	68.47	93.58	105.13

pitch group $F_o < 250$ Hz matches with the pitch range of the adults' training speech data.

From Table 6.2, it is to note that the changes in the speech recognition performances obtained by reducing the number of filters in the Mel filterbank are consistent with those reported in [123]. On analyzing the ASR performances of the children's test sets on both recognition tasks, it is noted that although a 11% relative improvement is obtained over baseline using 16 filters in the filterbank for MFCC feature computation for the children's test set CHts1 on the connected digit recognition task, no improvement in the overall performance is obtained by reducing the number of filters in the filterbank for MFCC feature computation for the children's test set PFts on the continuous speech recognition task. Also, significantly lesser relative improvements are obtained for all pitch groups with no improvement in case of less than 250 Hz pitch group on the continuous speech recognition task in comparison to those obtained on the connected digit recognition task. These inconsistent improvements across different pitch group children's speech signals are attributed to the fact that all filters used in the filterbank are constant-Q filters.

Modifying the number of filters in the Mel filterbank results in non-uniform modification of the

bandwidths of all filters in the filterbank as shown in Table 6.1. The ratio of increase in the bandwidth of the first filter on reducing the number of filters in the Mel filterbank from 21 down to 12 is 1.8 while for the twelfth filter it is 3.5. Greater increase is noted in the bandwidths for the higher order filters than for the lower order filters with reduction in the number of filters in Mel filterbank. This leads to change in the degree of smoothing of the speech spectrum over the entire frequency range. It is already noted in Section 4.2 that the pitch-dependent distortions occur mainly in the lower frequency range in the smoothed Mel spectral envelope. Therefore, modifying the bandwidths of all filters in the filterbank implicitly by reducing the number of filters in the filterbank causes under- or over-smoothing of the speech spectrum in some frequency ranges and thereby increasing the pitch-dependent distortions in the smoothed spectral envelope or leading to the loss of spectral information. This can be further understood by observing the smoothed spectra corresponding to MFCC features computed using different number of filters in the Mel filterbank for vowel /IY/ having pitch value of around 300 Hz shown in Figure 6.1. It is noted that though the degree of spectral smoothing increases as the number of filters in the filterbank reduce, the smoothing is effected over the complete speech spectrum irrespective of the location of the pitch-dependent distortions in the spectral envelope.

6.2.2 Selective Modification of Filter Bandwidths

As already noted in Section 4.2, the pitch-dependent distortions in the smoothed Mel spectral envelope appear predominantly below 1 kHz for high pitch signals. It is also argued that the possible cause for that is the insufficient smoothing of the pitch harmonics by the lower order filters in the Mel filterbank which have typical bandwidths of around 100 Hz. Based on this, it is hypothesized that for pitch normalization it would be advantageous to modify the bandwidths of only the lower order filters in the filterbank. Therefore, in this section, spectral smoothing as effected by selective modification of the bandwidths of only the selected filters in the filterbank is explored.

First, the bandwidth of all filters having the center frequencies lower than 1 kHz in the default Mel filterbank are modified to various constant bandwidth values as illustrated in Figure 6.2. Motivated by the possible variation in the average pitch values across the speakers, the bandwidth of all filters having the center frequencies below 1 kHz are modified to nine different constant bandwidth values in the range of 100-500 Hz in steps of 50 Hz. The impact of modifying the bandwidth of all filters having center frequency below 1 kHz to different values on the smoothed Mel spectrum for vowel /IY/ having average pitch value of around 300 Hz is shown in Figure 6.3. It is noted that greater

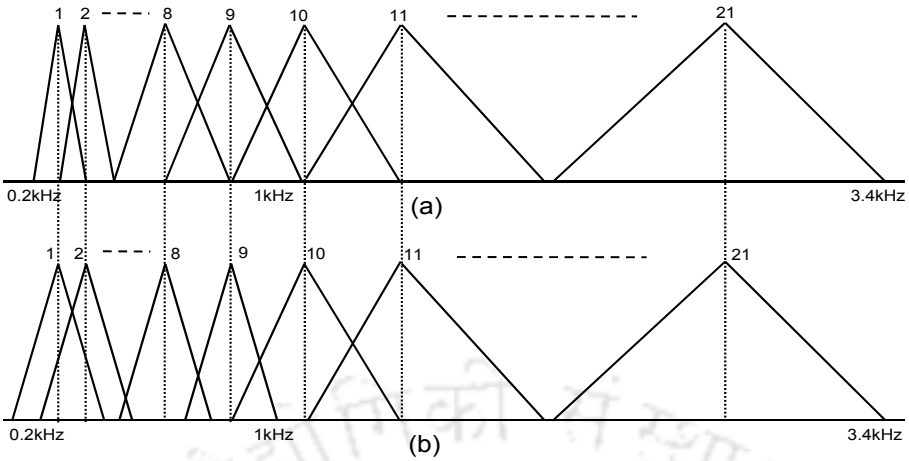


Figure 6.2: Structures of the Mel filterbank (a) Default (b) Modified. In the modified filterbank the bandwidth of all filters having center frequency below some particular frequency value (say 1 kHz) are modified to have a constant value whereas those of the other filters remain unchanged.

smoothing is obtained as the bandwidth of the filters is increased, thus, reducing the pitch-dependent distortions. On the other hand, expectedly, the pitch-dependent distortions are further enhanced when the bandwidth of the filters is lowered compared to the default filterbank case. In order to study the impact of such filterbank based spectral smoothing on the recognition performance, the bandwidth of the filters having center frequency below 1 kHz in the filterbank are modified during the MFCC

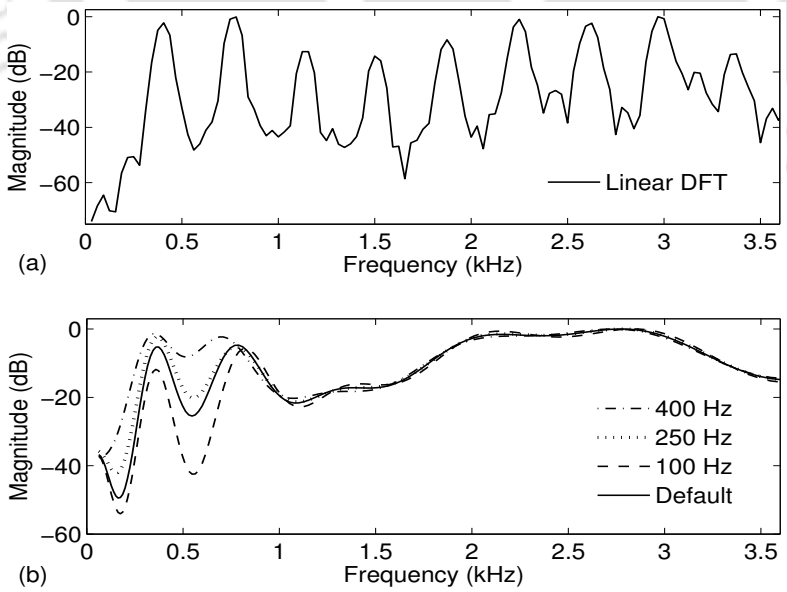


Figure 6.3: Plots for vowel /IY/ having pitch value of around 300 Hz (a) Linear DFT spectrum (b) Smoothed Mel spectra computed using various bandwidth values for all filters having center frequency below 1 kHz in Mel filterbank.

Table 6.3: Performances for children's test sets CHts1 (on connected digit recognition task) and PFts (on continuous speech recognition task) for various modified constant bandwidth (BW) values for all filters having center frequencies below 1 kHz in the filterbank along with their pitch group-wise breakup. Note that the best ASR performance for each pitch group corresponds to different choice of the bandwidth of filters considered.

BW (in Hz) of filters having center freq. below 1 kHz	% WER							
	CHts1				PFts			
	All F_o Values (7,772)	$F_o <$ 250 Hz (5,224)	250 Hz $\leq F_o <$ 300 Hz (2,346)	$F_o \geq$ 300 Hz (202)	All F_o Values (129)	$F_o <$ 250 Hz (63)	250 Hz $\leq F_o <$ 300 Hz (54)	$F_o \geq$ 300 Hz (12)
Default	11.37	6.54	17.47	39.03	56.34	33.05	77.25	102.69
100	16.56	9.09	26.54	53.41	64.67	37.27	89.34	118.83
150	12.40	7.03	19.17	43.25	60.59	34.95	84.55	106.85
200	10.37	6.46	15.29	32.84	55.34	32.71	74.93	103.91
250	**9.61**	**6.43**	13.60	28.25	**52.26**	**31.95**	**69.40**	98.04
300	9.69	6.74	**13.35**	27.26	53.86	35.41	69.45	95.35
350	10.12	7.18	13.86	26.64	57.39	39.32	74.28	**89.98**
400	10.36	7.44	14.12	26.15	61.12	43.58	77.64	92.18
450	10.57	7.75	14.21	**25.77**	65.17	44.15	84.25	105.87
500	10.81	8.03	14.36	26.02	68.17	47.26	85.74	115.65

feature computation of the test signals. The modified test features are then decoded with the acoustic models trained using the default MFCC features. On the connected digit recognition task, the speech recognition performances for the children's test set CHts1 corresponding to various modified constant bandwidth values for all filters having center frequency below 1 kHz along with their breakup for different pitch groups are given in Table 6.3. On the continuous speech recognition task, the speech recognition performances for the children's test set PFts corresponding to various modified constant bandwidth values for all filters having center frequency below 1 kHz along with their breakup for different pitch groups are also given in Table 6.3.

From Table 6.3, it is noted that the best recognition performance (in WER) of 9.61% (16% relative improvement over baseline) and 52.26% (7% relative improvement over baseline) is obtained using constant bandwidth value of 250 Hz for all filters having center frequency below 1 kHz in MFCC filterbank for the children's test set CHts1 (on connected digit recognition task) and the children's test

set PFts (on continuous speech recognition task), respectively. Further, on analyzing the performances of the children's test sets based on different pitch groups on both recognition tasks, it is observed that the modified constant bandwidth value for all filters having center frequency below 1 kHz in filterbank corresponding to the best performance for each pitch group increases with increasing pitch of the group. This supports our earlier observation that the magnitude of the pitch-dependent distortions increases with increasing pitch of the signals and thus, they require more smoothing with respect to the adults' speech trained acoustic models. As observed in Figure 6.3, the decrease in the performance corresponding to very low and very high values of the bandwidth of the filters are attributed to under- and over-smoothing of the resulting smoothed spectrum, respectively.

Motivated by the above noted different modified constant bandwidth values for all filters having center frequency below 1 kHz in filterbank corresponding to the best speech recognition performances for different pitch groups, the number of filters whose bandwidth is required to be modified for different pitch group signals of the children's test sets are also explored. The recognition performances for the children's test sets CHts1 and PFts using MFCC features computed using constant bandwidth values of 250 Hz and 300 Hz for different number of filters in the Mel filterbank are given in Table 6.4 and Table 6.5, respectively. It is to note that greater improvements are obtained on both recognition tasks by modifying the bandwidths of only the selected number of filters in the filterbank rather than for all filters.

On connected digit recognition task, the best ASR performances are obtained for the signals of the children's test set CHts1 having average pitch values of less than 250 Hz using constant bandwidth of 250 Hz for the first 3 filters in the Mel filterbank. On the other hand, for the signals of the children's test set CHts1 having average pitch values in the range from 250 Hz to 300 Hz and those having average pitch values of greater than or equal to 300 Hz, the best ASR performances are obtained using constant bandwidth of 300 Hz for the first 6 filters and the first 7 filters in the Mel filterbank, respectively. On continuous speech recognition task, the best recognition performances for the signals of the children's test set PFts having average pitch values of less than 250 Hz are obtained using constant bandwidth of 250 Hz for the first 11 filters in the Mel filterbank. On the other hand, using constant bandwidth of 300 Hz for the first 6 filters and the first 13 filters in the filterbank gives the best ASR performances for the signals of the children's test set PFts having average pitch values greater than or equal to 250 Hz but less than 300 Hz and those having average pitch values of greater than

Table 6.4: Performance for children's test set CHts1 (on connected digit recognition task) with modified constant bandwidth (BW) values of 250 Hz and 300 Hz for various number of filters in the filterbank along with its pitch group-wise breakup. Note that higher pitch groups require modification of the bandwidth of filters up to higher frequencies.

Indices of filters modified	% WER							
	Modified filter BW of 250 Hz				Modified filter BW of 300 Hz			
	All F_o Values	$F_o <$ 250 Hz	250 Hz $\leq F_o <$ 300 Hz	$F_o \geq$ 300 Hz	All F_o Values	$F_o <$ 250 Hz	250 Hz $\leq F_o <$ 300 Hz	$F_o \geq$ 300 Hz
Default	11.37	6.54	17.47	39.03	11.37	6.54	17.47	39.03
1-3	9.95	**6.26**	14.47	32.84	9.84	6.31	14.06	32.59
1-4	9.52	6.31	13.46	29.37	9.38	6.40	12.96	28.62
1-5	9.41	6.31	13.19	28.75	**9.26**	6.38	12.83	26.52
1-6	9.49	6.39	13.32	28.13	9.45	6.75	**12.70**	26.89
1-7	9.60	6.48	13.51	28.00	9.66	6.87	13.11	**26.39**
1-8	9.59	6.44	13.50	28.62	9.72	6.81	13.33	27.39
1-9	9.61	6.43	13.60	28.25	9.69	6.74	13.35	27.26
1-10	9.63	6.45	13.55	28.87	9.83	6.81	13.58	27.88
1-11	9.63	6.43	13.61	28.75	9.81	6.84	13.46	27.88
1-12	9.65	6.48	13.52	29.24	9.82	6.82	13.54	27.88
1-13	9.61	6.43	13.51	29.00	9.83	6.86	13.50	27.88
1-14	9.67	6.38	13.76	29.24	9.80	6.87	13.42	27.51
1-21	11.24	7.33	16.21	33.33	10.30	7.24	14.18	27.51

or equal to 300 Hz, respectively. The bandwidths of larger number of filters to be modified in Mel filterbank for the signals with the average pitch values of greater than 300 Hz in comparison to those having average pitch values in the range from 250 Hz to 300 Hz is again attributed to the appearance of the pitch-dependent distortions in the smoothed spectral envelope to a greater frequency range in case of high pitch signals due to greater number of filters having bandwidth smaller than that of the average pitch of the signals.

Therefore, greater and consistent improvements are obtained in the children's ASR performance for each pitch group by selectively modifying the bandwidth of only the selected lower order filters in the filterbank than those obtained either by explicitly modifying the bandwidth of all filters in the filterbank or by reducing the number of filters in the filterbank during MFCC feature computation

Table 6.5: Performance for children's test set PFts (on continuous speech recognition task) with modified constant bandwidth (BW) values of 250 Hz and 300 Hz for various number of filters in the filterbank along with its pitch group-wise breakup. Note that higher pitch groups require modification of the bandwidth of filters up to higher frequencies.

Indices of filters modified	% WER							
	Modified filter BW of 250 Hz				Modified filter BW of 300 Hz			
	All F_o Values	$F_o <$ 250 Hz	250 Hz $\leq F_o <$ 300 Hz	$F_o \geq$ 300 Hz	All F_o Values	$F_o <$ 250 Hz	250 Hz $\leq F_o <$ 300 Hz	$F_o \geq$ 300 Hz
Default	56.34	33.05	77.25	102.69	56.34	33.05	77.25	102.69
1-3	54.10	32.75	71.87	103.42	54.73	34.84	71.08	101.71
1-4	53.58	32.07	71.62	102.69	54.86	34.92	71.72	99.76
1-5	53.38	32.41	71.22	100.00	53.84	35.22	69.45	96.33
1-6	51.69	31.99	67.92	98.04	52.67	35.37	**66.83**	93.89
1-7	52.04	31.88	69.10	97.31	52.30	35.11	66.93	90.46
1-8	52.30	31.91	69.20	99.76	52.85	35.64	67.08	93.15
1-9	52.26	31.95	69.40	98.04	53.86	35.41	69.45	95.35
1-10	52.24	31.80	69.10	100.24	53.27	35.49	68.26	93.40
1-11	**51.39**	**31.50**	67.52	99.51	53.31	35.30	68.16	95.60
1-12	52.34	32.79	68.26	99.27	53.33	35.41	68.02	95.84
1-13	52.69	32.83	69.40	97.80	52.77	35.64	67.57	**89.73**
1-14	51.43	32.48	67.23	95.11	53.42	35.14	68.76	95.11
1-21	54.63	32.83	72.75	105.13	55.04	37.69	69.74	93.89

for children's speech. The merit of the selective bandwidth modification of filters in the filterbank for achieving pitch normalization is attributed to the selective smoothing of the speech spectrum such that only the required frequency range of the spectrum which is significantly affected by the pitch-dependent distortions is smoothed out and thus, avoiding the under- or over-smoothing of the speech spectrum over the entire frequency range. This fact is also supported by the observations made in [38] which studied the scaling of the bandwidths of all filters in Mel filterbank by factors ranging from 0.4 to 1.4 for MFCC feature computation and reported no significant improvements in the ASR performances for both adults' and children's speech.

6.3 Proposed Pitch Normalization Algorithm

In the previous section, significant improvements in the children's ASR performance are noted by selectively modifying the bandwidths of the filters in the filterbank for each pitch group children's test signals for MFCC feature extraction. These improvements have been attributed to the improved smoothing of the pitch harmonics in the speech spectrum by the modified filters of the filterbank. However, each pitch group consists of number of speech signals having different average pitch values. Therefore, it would be more appropriate to choose the modified constant bandwidth values for appropriate number of filters in the filterbank for pitch normalization for each speech signal independently.

Some correlation of the modified constant bandwidth values and the number of filters whose bandwidths need to be modified in the Mel filterbank with the average pitch of the speech signal for MFCC feature computation could be noted from the studies reported in the previous section. The speech signals belonging to the higher pitch group give best recognition performances using larger constant bandwidth values for larger number of filters in the Mel filterbank for MFCC feature computation. It is also noted that the modified constant bandwidth value for the filters giving the best ASR performance in case of each pitch group is not less than the minimum average pitch of that pitch group. This is obvious as to avoid the occurrence of the pitch-dependent distortions in the smoothed spectral envelope due to insufficient smoothing by the filters of MFCC filterbank, the bandwidth of all filters in the filterbank should at least be equal to the average pitch value of the signal. Following this, we propose an algorithm for adaptively modifying the bandwidths of the filters in the Mel filterbank for MFCC feature computation for each speech signal.

In the proposed algorithm, the bandwidths of all filters in the filterbank having bandwidth values lesser than the average pitch value of the speech signal is modified to constant bandwidth value equal to the average pitch of that speech signal during its MFCC feature computation. The pseudocode for the proposed algorithm for pitch adaptive modification of Mel filterbank for pitch normalization is given in Algorithm 1. The children's speech recognition performances obtained using this proposed algorithm for pitch normalization during MFCC feature computation of children's test speech signals on both connected digit recognition and continuous speech recognition tasks are given in Table 6.6. It is noted that significant relative improvements of 16% and 9% are obtained over the corresponding default baseline children's ASR performances by using the proposed pitch normalization algorithm on connected digit recognition task and continuous speech recognition task, respectively.

Algorithm 1 Proposed pitch adaptive Mel filterbank modification algorithm

Require: Average pitch of speech signal F_0 ; Number of filters in Mel filterbank N
Require: Default Mel filterbank (DC_1, DB_1),...,(DC_N, DB_N)
where DC_i is the center frequency and DB_i is the bandwidth of i^{th} filter

Initialize: Modified Mel filterbank $(MC \leftarrow (), MB \leftarrow ())$

```
for i = 1 → N do
  if DB_i < F_0 then
    MB_i ← F_0
  else
    MB_i ← DB_i
  end if
  MC_i ← DC_i
end for
```

Output modified Mel filterbank: (MC_1, MB_1),...,(MC_N, MB_N)

It is to note here that the improvements obtained with the proposed algorithm are comparable to those obtained in the previous section by searching the modified values for filter bandwidths. Also, the improvements are comparable to those obtained with the explicit pitch normalization approach with an additional advantage of reduced computational complexity in comparison to that in case of the explicit pitch normalization. For sake of comparison, the children's ASR performances obtained after explicit pitch normalization of children's speech on connected digit and continuous speech recognition tasks are also given in the last row in Table 6.6.

In literature, the children's matched ASR performance has been reported to be poorer than the adults' matched ASR performance [60]. This is attributed to the higher inter- and intra-speaker acoustic variability in case of children in comparison to that in case of adults [27,30]. In [63], increasing the number of filters with default bandwidth in the Mel filterbank has been reported for MFCC feature extraction for children's speech for improving their ASR performance on matched models. Motivated by these, we further explored the efficacy of the proposed modification in Mel filterbank for pitch normalization for children's ASR on children's speech trained models. A relative improvement of 32% is noted in the ASR performance of children's test set CHts2 on models trained on children's training set CHtr on continuous speech recognition task. It is to note that the improvement in the children's ASR performance on matched models with the proposed algorithm is greater than that reported in [63]. This is again attributed to the greater and selective smoothing of pitch harmonics in

Table 6.6: Performances for children's test sets CHts1 (on connected digit recognition task) and PFts (on continuous speech recognition task) using the default MFCC features and MFCC features computed using the proposed pitch normalization algorithm both with and without VTLN and CMLLR. The relative improvement (in %) for each case over its corresponding baseline is given in the parentheses.

Condition	% WER			
	CHts1		PFts	
	Default	Proposed	Default	Proposed
Baseline	11.37	9.50	56.34	51.25
with VTLN	2.95	2.35	26.78	24.99
(% Relative Gain)	(74)	(75)	(52)	(51)
with CMLLR	5.54	4.67	38.25	36.49
(% Relative Gain)	(51)	(51)	(32)	(29)
with Explicit Pitch Norm.	9.64		53.72	

the speech spectrum by the proposed algorithm. Thus, the proposed algorithm for pitch normalization is efficient for improving children's ASR not only on the adults' speech trained models but also on the children's speech trained models.

6.4 Combining Proposed Algorithm with VTLN and CMLLR

In this section, we explore whether the improvement obtained in the children's ASR performance on the adults' speech trained models using the proposed filterbank modification based pitch normalization algorithm during MFCC feature computation is additional to those obtained by the already existing techniques used for addressing the acoustic mismatch in context of ASR i.e, VTLN and CMLLR for children's speech recognition [30, 69] or not. The VTLN and CMLLR are performed on the test data using HTK as described in Chapter 2.

The ASR performances for the children's test sets CHts1 (on connected digit recognition task) and PFts (on continuous speech recognition task) on the adults' speech trained models both with and without VTLN on test data using the default MFCC features and the MFCC features computed using the proposed Mel filterbank based pitch normalization algorithm are also given in Table 6.6. It is noted that further improvements in the performances are obtained with relative gains of about 20% and 7% over corresponding default VTLN performances on using the proposed filterbank based pitch

normalization approach in conjunction with VTLN for the CHts1 and PFts test sets, respectively. Also, it is to note that the relative gain obtained with VTLN when performed in conjunction with the proposed filterbank based pitch normalization is comparable to that obtained by VTLN in the default case.

The modification of the constant-Q filterbank for VTLN purpose also results in small increase in the bandwidths of all filters of the Mel filterbank particularly in case of children's (high pitch) speech signals in HTK implementation. Thus, VTLN does provides some increase in the degree of smoothing of pitch harmonics in the speech spectrum for warp factors less than 1. However, the maximum change in the bandwidths of the lower order filters up to 1 kHz would be around 50 Hz only. This increase in the bandwidths of the filters is much small to result in any pitch normalization. This fact is also substantiated by the relative gain obtained with VTLN being comparable both with and without the combination of the proposed filterbank based pitch normalization algorithm.

Further, the children's speech recognition performances for the data sets CHts1 and PFts with CMLLR on test data using both default MFCC features and MFCC features computed using the proposed Mel filterbank based pitch normalization algorithm are also given in Table 6.6. It is noted that significant relative gains of 16% and 5% are obtained over corresponding default CMLLR performances by additionally performing the proposed filterbank based pitch normalization for children's test sets CHts1 and PFts, respectively. It is hypothesized that the further improvement with the proposed pitch normalization algorithm over the default CMLLR performance is due to the significant additional reduction in the acoustic mismatch due to non-linear pitch differences besides that addressed by CMLLR which is limited to linear transformations only. Also, the relative gain obtained with CMLLR when performed in conjunction with the proposed filterbank based pitch normalization is comparable to that obtained with CMLLR in the default case.

Therefore, the significant improvement obtained in the children's ASR performance with the proposed pitch normalization algorithm is additive to those obtained with the existing VTLN and CMLLR techniques with relative gains obtained with VTLN and CMLLR being comparable to those obtained in their corresponding default cases.

6.5 Summary

In this chapter, the work is motivated by our previous study on the effect of pitch differences across speech signals on MFCC features and the improvements obtained in the children's speech recognition performance with explicit ML-based pitch normalization. The work presented in this chapter and the observations made in this study are summarized below.

- An automatic algorithm is proposed for pitch normalization during MFCC feature extraction to avoid the computationally expensive ML-based explicit pitch normalization approach for children's speech.
- The proposed algorithm normalizes the pitch differences across speech signals through improved smoothing of the pitch harmonics in the speech spectrum. The bandwidths of the filters in the Mel filterbank are selectively modified during MFCC feature extraction for each children's test speech signal based on the average pitch of the test signal.
- Significant improvements are obtained in the children's ASR performances on the adults' speech trained models using the proposed pitch normalization algorithm comparable to those obtained with explicit ML-based pitch normalization technique on both connected digit recognition and continuous speech recognition tasks.
- Proposed algorithm for utterance-specific modification of the bandwidths of only selected filters in the Mel filterbank is noted to give larger improvement than those obtained either by explicitly modifying the bandwidths of all filters or by reducing the number of filters in the filterbank.
- Additional improvement is obtained in the children's ASR performance by using the proposed pitch normalization algorithm over those obtained with VTLN and CMLLR in default case. Also, the relative gains obtained with VTLN and CMLLR when applied in combination with the proposed pitch normalization algorithm are found to be comparable to those obtained with VTLN and CMLLR in the corresponding default cases.

7

Pitch Mismatch Reduction by Cepstral Truncation

Contents

7.1 Introduction 112
7.2 Truncation of MFCC Features for Children's ASR 112
7.3 Role of MFCC Feature Truncation in Pitch Mismatch Reduction 118
7.4 Adaptive MFCC Feature Truncation for Pitch Mismatch Reduction . . 122
7.5 Combining Proposed Algorithm with VTLN and CMLLR 133
7.6 Summary 135

7.1 Introduction

In Section 4.3, the pitch-dependent distortions appearing in the Mel spectral envelope of the high pitch signals have been found to increase the dynamic range of the higher order coefficients of 13-D truncated MFCC ($C_0 - C_{12}$) features. This in turn causes the variances of those higher order MFCCs to increase with increase in the pitch of the signals as noted in Section 3.2.1. It is already known that children's speech have higher pitch frequencies in comparison to those of the adults' speech [27]. Therefore, higher order coefficients of MFCC features corresponding to children's speech would have significantly higher variances in comparison to those in case of the adults' speech.

Motivated by these, in this chapter, the truncation of MFCC features is explored for children's ASR on adults' speech trained models. The role of MFCC feature truncation in reducing the pitch mismatch between the adults' and the children's speech is then explored. Based on the correlation observed between the length of the base MFCC features for a test signal and the degree of acoustic mismatch with respect to speech recognition acoustic models, an automatic algorithm is proposed for choosing utterance-specific appropriate truncation of MFCC features for children's speech recognition on adults' speech trained models.

The outline of this chapter is as follows: In Section 7.2, truncation of MFCC features is explored for children's speech recognition on adults' speech trained models. The role of MFCC feature truncation in addressing the pitch mismatch between the adults' and the children's speech in context of children's ASR on adults' speech trained models is studied in Section 7.3. In Section 7.4, an automatic algorithm for utterance-specific MFCC truncation for test signals is proposed. VTLN and CMLLR techniques are explored in combination with the proposed algorithm for ASR in Section 7.5. Finally, the summary of the work presented in this chapter is given in Section 7.6.

7.2 Truncation of MFCC Features for Children's ASR

To address the increase in the variances of the higher order coefficients of MFCC features with increase in the pitch of the signals, in this study, the effect of exclusion of the higher order coefficients from the default 13-D base MFCC features ($C_0 - C_{12}$) is explored on the children's ASR performance. The base MFCC features for the children's test speech are truncated from 13 down to 3 in steps of 1. For recognition, the various truncated base MFCC features are also appended with their corresponding first and second order temporal derivatives. The extended truncated test features are then decoded

using the corresponding dimensionality models. For optimal ASR evaluations, the acoustic models are required to be re-estimated to match the dimensionality of the truncated test features. However, in order to avoid the large computational complexity and the need of a large storage space, we explore using the suboptimal truncated models extracted from the full 39-D baseline models for ASR evaluations. On the connected digit recognition task, a difference of only 0.3% is noted in the children's mismatched ASR performances obtained using the re-trained models and the truncated models corresponding to the 3-D base MFCC features. Therefore, in this work, though all ASR experiments on the connected digit recognition task are done using the re-estimated reduced dimensionality models but, on the continuous speech recognition task the ASR performances are evaluated using the truncated models extracted from the 39-D baseline models.

The recognition performances for the children's test sets CHts1 and PFts on their respective systems for different dimensions of the truncated test features on corresponding adults' speech trained models with matching feature dimensions are given in Table 7.1. It is noted that the recognition performance for the children's test set improves consistently with MFCC truncation on both connected digit and continuous speech recognition tasks. The best relative improvement of about 54% is obtained over the baseline for the CHts1 test set for 12-dimensional MFCC features (include 4-dimensional base features and their first and second order temporal derivatives) while the best relative improvement of about 38% is obtained over the baseline for the PFts test set using 18-dimensional MFCC features (include 6-dimensional base features and their first and second order temporal derivatives). Thus, greater degree of cepstral truncation of default base MFCC features helps children's speech recognition on the adults' speech trained models.

It seemed bit surprising that such an improved children's ASR performance could be obtained on the adults' speech trained models with base MFCC feature length of as low as 4 (on connected digit recognition task) and 6 (on continuous speech recognition task) in comparison to default base MFCC feature length of 13. So, to explore the reason for such behavior, the contribution of each of the coefficients of the default 39-D MFCC features ($C_0 - C_{12}$ base MFCC features and their first and second order temporal derivatives) to the children's ASR performance is evaluated. For sake of ease, the study is done on a connected digit recognition task.

For determining the contribution of each of the coefficients of the default 39-D MFCC features to the children's ASR performance, separate acoustic model sets are derived from the 39-D baseline

Table 7.1: Performances for children's test set using MFCC features consisting of varying base feature dimensions on both connected digit recognition and continuous speech recognition tasks. For recognition, the various truncated base MFCC features are also appended with their corresponding first and second order temporal derivatives.

Recognition Task	% WER										
	Base MFCC Features										
	Default ($C_0 - C_{12}$)	$C_0 - C_{11}$	$C_0 - C_{10}$	$C_0 - C_9$	$C_0 - C_8$	$C_0 - C_7$	$C_0 - C_6$	$C_0 - C_5$	$C_0 - C_4$	$C_0 - C_3$	$C_0 - C_2$
Connected Digit	11.37	11.20	11.38	10.71	9.03	7.80	6.77	6.21	6.03	**5.21**	5.47
Continuous Speech	56.34	52.10	48.39	44.72	42.19	39.89	39.02	**35.13**	38.25	41.50	40.62

Table 7.2: Performances (in descending order) of each of the coefficients of the 39-D default MFCC features for children's test set CHts1 on models trained on adults' speech data set ADtr.

Rank	Coefficient	% WER	Rank	Coefficient	% WER	Rank	Coefficient	% WER
1	ΔC_0	**63.12**	14	C_{10}	88.16	27	ΔC_6	96.29
2	C_1	67.41	15	C_6	88.51	28	ΔC_{11}	96.33
3	C_0	69.93	16	C_5	89.17	29	ΔC_9	96.49
4	ΔC_1	72.98	17	C_8	90.24	30	ΔC_{10}	96.86
5	ΔC_2	73.16	18	C_9	91.49	31	$\Delta\Delta C_5$	97.24
6	$\Delta\Delta C_0$	74.47	19	C_{11}	92.07	32	ΔC_7	97.41
7	C_2	76.25	20	$\Delta\Delta C_3$	92.24	33	$\Delta\Delta C_6$	97.47
8	$\Delta\Delta C_2$	84.50	21	C_7	92.39	34	$\Delta\Delta C_{12}$	97.74
9	C_3	85.43	22	C_{12}	92.57	35	$\Delta\Delta C_9$	97.80
10	C_4	85.66	23	$\Delta\Delta C_4$	93.71	36	$\Delta\Delta C_{11}$	97.84
11	$\Delta\Delta C_1$	86.38	24	ΔC_5	94.95	37	$\Delta\Delta C_8$	98.03
12	ΔC_3	87.87	25	ΔC_{12}	95.70	38	$\Delta\Delta C_7$	98.23
13	ΔC_4	88.12	26	ΔC_8	96.27	39	$\Delta\Delta C_{10}$	**98.28**

acoustic model set trained on ADtr data set for each MFCC coefficient. The model set corresponding to a MFCC coefficient is derived by selecting the dimension corresponding to that particular coefficient from all mean and variance parameters of the 39-D baseline acoustic model set while keeping all the mixture weights and the transition probabilities same as default. The recognition performance of CHts1 data set is computed on each of those separate acoustic models with matching dimension of MFCC test features. This procedure is repeated separately for all coefficients of the default 39-D MFCC features and the recognition performances in descending order on the children's test set CHts1 computed using each of the individual coefficients of the 39-D MFCC features are given in Table 7.2. It is to note that some coefficients of the default 39-D MFCC features seem to have no significant role in children's ASR on adults' speech trained models.

Further, the performances with different MFCC features obtained by taking top performing coefficients as noted from Table 7.2 are also computed and are given in Table 7.3 along with those obtained with truncated MFCC features of matching length for ease of comparison. It is noted that the best relative improvement of 53% is obtained in the children's ASR performance using rank-ordered MFCC features of 12 dimensions for both training and test speech data. It is also noted that comparable

Table 7.3: Performances for children's test set CHts1 on models trained on adults' speech data set ADtr using MFCC feature vectors of different dimensions selecting the top d coefficients from the rank-ordered list given in Table 7.2 based on the contribution of each of the coefficients of 39-D MFCC feature vector to the speech recognition performances for the children's test set CHts1 on models trained on adults' speech data set ADtr, referred to as 'Rank Based Feature Selection'. For ease of comparison, the children's ASR performances for CHts1 test set corresponding to MFCC features truncated to different dimensions obtained in Table 7.1 are also given.

Method	% WER on Connected Digit Recognition Task												
	MFCC Feature Length												
	3	6	9	12	15	18	21	24	27	30	33	36	39
Rank Based Feature Selection	24.51	10.24	5.99	**5.35**	6.03	7.81	8.64	8.73	9.32	9.93	10.41	10.74	11.37
Cepstral Truncation Based Feature Selection	-	-	5.47	**5.21**	6.03	6.21	6.77	7.80	9.03	10.71	11.38	11.20	11.37

Table 7.4: Rank ordering of each of the coefficients of the 39-D MFCC features based on their ASR performances for children's test set CHts1 on models trained on adults' speech data set ADtr within the base, the Δ and the ΔΔ feature streams.

Base MFCC		Δ MFCC		ΔΔ MFCC	
Rank	Coefficient	Rank	Coefficient	Rank	Coefficient
1	$\mathbf{C_1}$	1	$\mathbf{\Delta C_0}$	1	$\mathbf{\Delta\Delta C_0}$
2	$\mathbf{C_0}$	2	$\mathbf{\Delta C_1}$	2	$\mathbf{\Delta\Delta C_2}$
3	$\mathbf{C_2}$	3	$\mathbf{\Delta C_2}$	3	$\mathbf{\Delta\Delta C_1}$
4	$\mathbf{C_3}$	4	$\mathbf{\Delta C_3}$	4	$\mathbf{\Delta\Delta C_3}$
5	$\mathbf{C_4}$	5	$\mathbf{\Delta C_4}$	5	$\mathbf{\Delta\Delta C_4}$
6	C_{10}	6	ΔC_5	6	$\Delta\Delta C_5$
7	C_6	7	ΔC_{12}	7	$\Delta\Delta C_6$
8	C_5	8	ΔC_8	8	$\Delta\Delta C_{12}$
9	C_8	9	ΔC_6	9	$\Delta\Delta C_9$
10	C_9	10	ΔC_{11}	10	$\Delta\Delta C_{11}$
11	C_{11}	11	ΔC_9	11	$\Delta\Delta C_8$
12	C_7	12	ΔC_{10}	12	$\Delta\Delta C_7$
13	C_{12}	13	ΔC_7	13	$\Delta\Delta C_{10}$

improvements are obtained in the children's ASR performance along with similar reduction in the overall dimensionality of MFCC features from both rank based low-dimensional MFCC features and low-dimensional MFCC features obtained through cepstral truncation. However, it is difficult to explain the spectral meaning of such arbitrary selection of coefficients in case of rank based MFCC features.

On observing the top 12 performing coefficients of 39-D MFCC features from Table 7.2, it is noted that the rank based 12-D MFCC features comprise mainly the coefficients up to C_3 and their first and second order derivatives. Further, on observing the rank ordering of all coefficients among the base, the Δ and the ΔΔ MFCC feature streams in Table 7.4 based on their performances as noted in Table 7.2, it is noted that the higher contributions to recognition performance in each of the three streams of MFCC features have come from the lower order coefficients. This justifies the comparable best improvement obtained in the children's ASR performance for CHts1 test set using truncated base MFCC features of 4 dimensions ($C_0 - C_3$).

7.3 Role of MFCC Feature Truncation in Pitch Mismatch Reduction

In previous section, it is noted that exclusion of higher order coefficients of MFCC features results in large improvements for children's speech recognition on adults' speech trained models. The children's speech have much higher pitch frequencies in comparison to those of the adults' speech which increase the dynamic range of the higher order coefficients of MFCC features as already shown in Chapter 4. Thus, the reason of the large improvements obtained with truncation of the higher order coefficients of MFCC features for children's ASR is obvious but it would be beneficial to understand the role of MFCC feature truncation in pitch mismatch reduction in more detail.

In HMM-based speech recognition system, the likelihood estimation for recognition purpose essentially involves the use of the MD [102] measure. Obviously, the larger MDs lead to poorer speech recognition performance. The MD for a MFCC test feature vector with respect to a low pitch speech trained model is computed using Eqn. 7.1:

$$MD(\mathbf{x}, \mu_L) = \sqrt{(\mathbf{x} - \mu_L)^T \mathbf{\Sigma}_L^{-1} (\mathbf{x} - \mu_L)} \tag{7.1}$$

where, $\mathbf{x}$ denotes the MFCC test feature vector, μ_L represents the mean and $\mathbf{\Sigma}_L$ represents the covariance of the low pitch speech trained model. Let, $\mathbf{x}_H$ denote MFCC test feature vector of high pitch speech signal and $\mathbf{x}_L$ denote MFCC test feature vector of low pitch speech signal. Let 'i' denote the indices of the higher order coefficients of base MFCC features and their corresponding Δ and $\Delta\Delta$ coefficients. Then, based on the observations made in Section 4.3,

$$x_{H_i} >> \mu_{L_i} \quad \text{while} \quad x_{L_i} \approx \mu_{L_i} \tag{7.2}$$

$$\Rightarrow \mathbf{x}_H - \mu_L > \mathbf{x}_L - \mu_L \tag{7.3}$$

$$\Rightarrow MD_H > MD_L \tag{7.4}$$

Therefore, we hypothesize that the higher order MFCC would have much higher distances than those for the lower order MFCC for signals having high pitch values with respect to the models trained with signals having low pitch values which have comparatively low variances across all coefficients.

In order to explore the relative contribution of the lower and the higher order coefficients of MFCC ($C_1 - C_{12}$) features in the distance measure in case of both low and high pitch signals, the variances of squared MD of 13-D base MFCC feature vectors of signals belonging to different pitch ranges are

measured for different order of truncation of MFCC features. For this study, the TIMIT data is used and nearly 2000 frames (central steady-state portions) of 7 different vowels from speech signals belonging to 75-100 Hz, 100-125 Hz and 200-250 Hz pitch ranges are extracted. The average pitch (F_o) of the signals is estimated using the ESPS tool available in the Wavesurfer software package [88] as described in Section 2.2. The variances of the squared MD of the MFCC feature vectors of the 'low' (100-125 Hz) and the 'high' (200-250 Hz) pitch group signals from the distribution of MFCC features of 75-100 Hz pitch group signals corresponding to different truncations of MFCC features are measured for different vowels and those for few representative vowels are given in Table 7.5. The MD is computed for the MFCC feature vectors of all signals of both pitch groups using Eqn. 7.1 where $\mathbf{x}$ is the MFCC feature vector whose distance is to be computed. μ_L is the mean and $\mathbf{\Sigma}_L$ is the diagonal-covariance of the distribution of MFCC features of 75-100 Hz pitch group signals. The table also shows the ratio of the variances of the squared MD of the feature vectors of the 'low' and 'high' pitch group signals for different order of truncation of MFCC features.

From Table 7.5, it is noted that, similar to our observation in Section 3.2.1, for MFCC features of all feature lengths the feature vectors of high pitch group signals have larger variances of squared MD with respect to the 75-100 Hz pitch group distribution in comparison to those for the low pitch group signals. As the degree of truncation of MFCC features increases the variance of squared MD of feature vectors of both pitch groups decreases for all vowels. However, the decrement in the variances of squared MD is more for high pitch group signals than the low pitch group signals in all cases.

Also, it is to note that greater decrease in the variances of squared MD is observed when the higher order MFCCs (beyond C_4) are excluded from the feature vector in comparison to the exclusion of the lower order MFCC (up to C_4) for both pitch groups. Observing the ratio of the variances of squared MD of feature vectors of the high and the low pitch groups, it is noted that with increase in truncation of higher order MFCCs the ratio of the variances of squared MD of feature vectors of both groups decreases significantly with the most decrease on exclusion of the coefficients $C_{10} - C_{12}$. This indicates that with increase in truncation of MFCC features the feature vectors of high pitch group signals come closer to the distribution of MFCC features of the 75-100 Hz pitch group signals to an extent similar to those of the low pitch group but more when the higher order MFCC are truncated by reducing the pitch mismatch. This supports our earlier hypothesis that the higher order MFCCs have much higher distances in comparison to those for the lower order MFCC for high pitch signals compared to those

Table 7.5: Variances of squared Mahalanobis distances of MFCC feature vectors of the 'low' (100-125 Hz) and the 'high' (200-250 Hz) pitch group signals from the distribution of MFCC features of 75-100 Hz pitch group signals with different truncations of MFCC features for different vowels.

Base MFCC	Vowel /AE/			Vowel /IY/			Vowel /UW/		
Features	F_o =100-125 Hz	F_o =200-250 Hz	Ratio	F_o =100-125 Hz	F_o =200-250 Hz	Ratio	F_o =100-125 Hz	F_o =200-250 Hz	Ratio
Default	60.2	2553.1	**42.4**	85.3	1735.4	**20.3**	114.4	1826.5	**16.0**
$\mathbf{C_1 - C_{11}}$	49.4	1191.6	**24.1**	75.8	1076.8	**14.2**	89.8	968.7	**10.8**
$\mathbf{C_1 - C_{10}}$	41.1	459.8	**11.2**	60.9	524.3	**8.6**	66.2	339.7	**5.1**
$\mathbf{C_1 - C_9}$	34.2	195.1	**5.7**	50.5	254.8	**5.0**	60.1	236.6	**3.9**
$C_1 - C_8$	29.0	153.8	5.3	41.7	101.1	2.4	36.4	78.3	2.2
$C_1 - C_7$	23.7	111.4	4.7	31.3	71.5	2.3	22.4	54.5	2.4
$C_1 - C_6$	20.2	55.4	2.7	22.8	53.6	2.4	17.1	31.6	1.8
$C_1 - C_5$	14.2	53.2	3.7	16.2	47.6	2.9	13.7	31.1	2.3
$C_1 - C_4$	10.1	30.5	3.0	12.0	40.9	3.4	9.5	21.2	2.2
$C_1 - C_3$	7.9	27.0	3.4	8.4	30.4	3.6	6.7	13.4	2.0
$\mathbf{C_1 - C_2}$	5.0	13.8	**2.8**	4.5	5.7	**1.3**	3.6	4.8	**1.3**
$\mathbf{C_{11} - C_{12}}$	6.1	429.6	**70.4**	7.1	624.5	**88.0**	10.2	414.3	**40.6**
$C_{10} - C_{12}$	9.1	662.6	72.8	12.3	1017.4	82.7	13.0	544.2	41.9
$C_9 - C_{12}$	11.4	718.5	63.0	17.0	1232.2	72.5	21.1	1156.0	54.8
$C_8 - C_{12}$	15.4	806.0	52.3	23.0	1217.3	52.9	21.7	1119.2	51.6
$C_7 - C_{12}$	19.4	1196.5	61.7	26.7	1323.8	49.6	31.8	1199.9	37.7
$C_6 - C_{12}$	27.5	1901.4	69.1	32.1	1309.0	40.8	35.6	1218.1	34.2
$C_5 - C_{12}$	31.9	2258.7	70.8	42.2	1360.9	32.2	43.2	1186.3	27.5
$C_4 - C_{12}$	38.3	2492.4	65.1	52.2	1450.9	27.8	55.5	1284.3	23.1
$C_3 - C_{12}$	44.0	2394.1	54.4	65.4	1538.4	23.5	80.2	1503.8	18.8
$C_2 - C_{12}$	54.9	2411.5	43.9	78.4	1626.3	20.7	93.9	1518.2	16.2

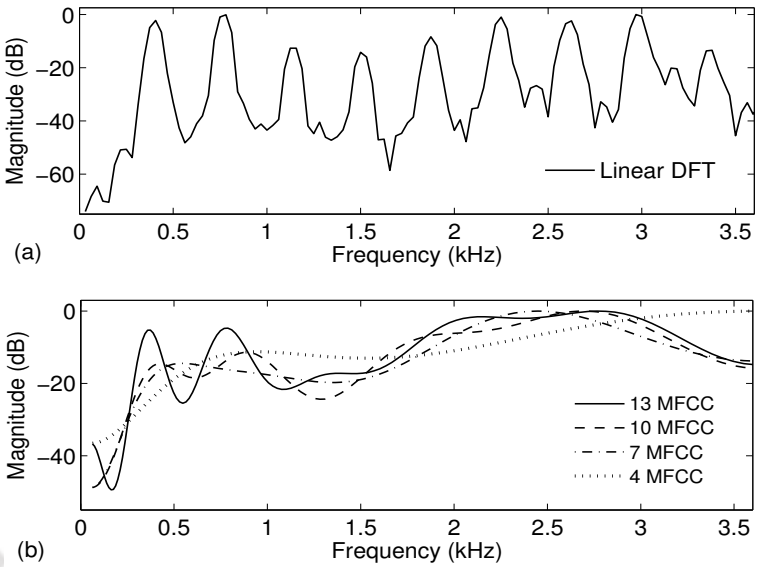


Figure 7.1: Plots for vowel /IY/ having pitch value of around 300 Hz (a) Linear DFT spectrum (b) Smoothed Mel spectra corresponding to the base MFCC features of different dimensions.

corresponding to the low pitch signals. Since children's speech have significantly higher pitch values than adults' speech, the exclusion of the higher order coefficients from the default 13-D base MFCC features ($C_0 - C_{12}$) helps in reducing the distance between the children's test features and the adults' speech trained models by reducing their pitch mismatch.

To illustrate the effect of varied truncation of MFCC features on their corresponding spectra, the plots of the smoothed spectra corresponding to various truncated base feature lengths including C_0 for vowel /IY/ having pitch value of around 300 Hz are shown in Figure 7.1. It is to note that with increased truncation, the pitch-dependent distortions in the smoothed Mel spectral envelope are better smoothed out. This explains the reduction in the pitch mismatch by increased truncation of higher order MFCCs. However, it is to note here that with greater truncation of higher order MFCCs the spectral peaks (formants) also start getting smoothed out. This truncation is applied to both the test features as well as the mean and variance parameters of the acoustic model. This corresponds to similar smoothing being applied to the adults' speech spectra as well. This results in reduction in the spectral mismatch between the adults' and children's speech spectra due to vocal tract length (VTL) differences. Thus, with the increased MFCC feature truncation for children's ASR on adults' speech trained models has the potential to reduce the acoustic mismatch between the adults' and the children's speech arising due to the pitch differences, the VTL differences and any other sources of acoustic mismatch which induce fast varying changes in the speech spectrum.

7.4 Adaptive MFCC Feature Truncation for Pitch Mismatch Reduction

As already noted in previous section that the increased truncation of MFCC features for children's test sets helps in reducing their acoustic mismatch due to pitch and VTL differences with respect to the adults' speech trained models. However, for each test speech signal the degree of acoustic mismatch with respect to the models is different. Therefore, in this section, we first explore the correlation between the appropriate MFCC feature truncation for a test signal and its degree of acoustic mismatch with respect to the speech recognition models. Following the observed correlation, algorithms are proposed for adaptive truncation of MFCC features of test signals for reducing their pitch mismatch with respect to the ASR models.

7.4.1 Correlation between MFCC Feature Truncation and VTLN Warp Factor

Although the acoustic mismatch of a test signal correlates with its likelihood with respect to the ASR models, in case of the truncated test features the likelihoods with respect to the models of matching dimensions would obviously be monotonically increasing with increasing truncation. Thus, the ML criterion can not be simply used to determine the appropriate truncation for a test feature required for its recognition unless the likelihoods are appropriately penalized. But those penalties for different truncations of MFCC features are to be determined empirically as deriving them analytically for HMM-based acoustic modeling is comparatively difficult. Further, the increased MFCC feature truncation reduces the acoustic mismatch mainly due to the pitch differences and the differences in the formant frequencies. Among these two factors, the format frequencies are already noted to be the foremost source of acoustic mismatch between adults' and children's speech. In addition to that, VTLN provides an easy quantification of acoustic mismatch in terms of the frequency warp factor. Therefore, we intend to explore the correlation between the degree of cepstral truncation of a MFCC test feature and its degree of acoustic mismatch with respect to speech recognition models as assessed by its ML-based VTLN warp factor estimate.

The VTLN warp factors for all speech signals of both children's test sets CHts1 and PFts are estimated with respect to the corresponding 39-D baseline adults' speech trained models by doing a ML grid search among 13 frequency warp factors (α) ranging from 0.88-1.12 in steps of 0.02. The speech signals of both children's test sets are then divided into different groups based on their ML-based

optimal VTLN warp factors. The VTLN warp factor-wise recognition performances of the children's test sets CHts1 and PFts for different dimensions of the truncated test features on corresponding adults' speech trained models with matching feature dimensions are given in Table 7.6 and Table 7.7, respectively. It is to note that there are not sufficient number of signals corresponding to each of the values of the VTLN warp factor in the children's speech test sets. So, to study the correlation between MFCC truncation and degree of acoustic mismatch for speech signals of other VTLN warp factors as well the recognition performances for the matched adults' speech test sets are also evaluated. The matched ASR performances of the adults' speech test sets ADts and CAMts with varying truncation of MFCC features along with their VTLN warp factor-wise performances are given in Table 7.8 and Table 7.9, respectively. It is noted that no significant change in the ASR performance occurs up to a length of 11 for base MFCC features for ADts test set while it slightly improves over baseline with increased truncation of MFCC features up to the length of 12 for base MFCC features for CAMts data set. The comparatively less feature truncation and small improvement for ADts and CAMts test sets are attributed to the lesser degree of gross acoustic mismatch in case of adults' test speech data with respect to the adults' speech trained models unlike in case of children's test speech data. Observing the VTLN warp factor-wise performances of significantly large signal groups from both children's and adults' speech test sets, it is noted that different truncations of base MFCC features are required for signals with different degree of VTL differences with greater truncation chosen for signals having greater VTL differences with respect to the adults' speech trained models. However, it is also worth noting that lesser degree of cepstral truncation is required for the adults' test speech signals in comparison to those required for the children's test speech signals having same VTLN warp factor values w.r.t. the adults' speech trained models.

7.4.2 Proposed Algorithm for Adults' Speech Trained ASR Models

Based on the observed correlation between the appropriate MFCC test feature truncation and its degree of acoustic mismatch, further, we propose an automatic algorithm for doing an utterance-specific truncation of default base MFCC features for test signals depending upon their acoustic mismatch with respect to the adults' speech trained models. To show the generality of the proposed algorithm, the ASR performances are evaluated using the same proposed algorithm for both the children's as well as the adults' test sets on both the connected digit recognition and continuous speech recognition tasks.

Table 7.6: Performance for children's test set CHts1 on models trained on adults' speech data set ADtr 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 parentheses gives the number of utterances corresponding to that VTLN warp factor.

MFCC Base Feature	% WER												
		VTLN Warp Factor Values											
	All	0.88	0.90	0.92	0.94	0.96	0.98	1.00	1.02	1.04	1.06	1.08	1.10
Length	(7,772)	(1384)	(4782)	(740)	(278)	(366)	(88)	(112)	(9)	(8)	(1)	(2)	(2)
Default ($C_0 - C_{12}$)	11.37	16.24	10.89	8.00	14.45	6.36	4.02	12.37	22.22	6.67	**0.00**	100.00	100.00
$C_0 - C_{11}$	11.20	16.27	10.83	7.37	13.01	6.00	3.57	10.60	22.22	6.67	0.00	100.00	**50.00**
$C_0 - C_{10}$	11.38	16.12	11.10	7.78	12.86	6.00	3.57	10.60	16.67	6.67	0.00	50.00	100.00
$C_0 - C_9$	10.71	14.58	10.57	6.91	12.57	5.91	3.57	11.66	16.67	6.67	0.00	50.00	100.00
$C_0 - C_8$	9.03	12.25	8.96	6.14	10.26	4.39	**2.23**	9.19	22.22	6.67	0.00	**0.00**	50.00
$C_0 - C_7$	7.80	10.49	7.71	5.41	9.68	4.03	2.68	7.42	11.11	**0.00**	0.00	0.00	50.00
$C_0 - C_6$	6.77	9.68	6.62	4.32	8.24	3.31	2.23	6.71	5.56	0.00	0.00	50.00	50.00
$C_0 - C_5$	6.21	8.89	6.01	**4.18**	7.51	**3.13**	2.68	**6.01**	5.56	6.67	0.00	50.00	50.00
$C_0 - C_4$	6.03	8.50	5.77	4.23	7.80	3.58	2.23	7.77	5.56	6.67	0.00	50.00	50.00
$C_0 - C_3$	**5.21**	**6.37**	**5.10**	4.41	**6.07**	3.58	3.57	6.36	**0.00**	6.67	0.00	50.00	50.00
$C_0 - C_2$	5.47	6.64	5.36	4.55	7.80	3.13	3.12	6.71	0.00	6.67	0.00	100.00	100.00

Table 7.7: Performance for children's test set PFts on models trained on adults' speech data set CAMtr for various truncations of base MFCC features along with their VTLN warp factor-wise breakup.

MFCC Base Feature Length	% WER								
		VTLN Warp Factor Values							
	All (129)	0.88 (91)	0.90 (24)	0.92 (7)	0.94 (1)	0.96 (2)	0.98 (2)	1.06 (1)	1.08 (1)
Default ($C_0 - C_{12}$)	56.34	64.07	40.97	26.32	25.00	**8.42**	94.44	42.86	5.13
$C_0 - C_{11}$	52.10	58.70	38.95	23.08	25.00	10.53	93.52	42.86	5.13
$C_0 - C_{10}$	48.39	54.76	34.21	22.67	25.00	**8.42**	93.52	42.86	5.13
$C_0 - C_9$	44.72	50.21	31.48	22.67	25.00	9.47	98.15	28.57	5.13
$C_0 - C_8$	42.19	47.27	28.96	22.27	25.00	**8.42**	99.07	30.95	5.13
$C_0 - C_7$	39.89	44.87	28.15	16.60	25.00	**8.42**	90.74	26.19	5.13
$C_0 - C_6$	39.02	43.99	27.35	16.19	13.64	**8.42**	93.52	21.43	5.13
$C_0 - C_5$	**35.13**	**39.56**	**23.71**	**14.98**	13.64	9.47	88.89	26.19	2.56
$C_0 - C_4$	38.25	42.25	29.26	16.60	22.73	9.47	90.74	23.81	2.56
$C_0 - C_3$	41.50	45.62	35.22	15.38	**9.09**	12.63	86.11	21.43	2.56
$C_0 - C_2$	40.62	44.22	33.70	21.05	**9.09**	25.26	**82.41**	**16.67**	**0.00**

It has already been noted that the adults' speech (matched) and the children's speech (mismatched) have different degrees of acoustic mismatches with respect to the adults' speech trained models, and therefore, choose different truncations of base MFCC features for improved ASR i.e., lesser truncation for matched adults' speech while greater truncation for mismatched children's speech on adults' speech trained models. Thus, it can be concluded that, in general, matched test speech would require lesser truncation in comparison to the mismatched test speech for recognition. Also, it has been observed that in case of both matched and mismatched ASR the appropriate base MFCC feature truncation increases with increase in the degree of acoustic mismatch between the test speech and the adults' speech trained models. Therefore, following these observations, for the sake of generality, we propose a piece-wise linear relationship between the length of base MFCC features and the VTLN warp factor ($\hat{\alpha}$) for both matched and mismatched test speech signals as shown graphically in Figure 7.2.

Table 7.8: Performance for adults' test set ADts on models trained on adults' speech data set ADtr for various truncations of base MFCC features along with their VTLN warp factor-wise breakup.

MFCC Base Feature Length	% WER													
		VTLN Warp Factor Values												
	All (3,303)	0.88 (46)	0.90 (646)	0.92 (424)	0.94 (301)	0.96 (460)	0.98 (166)	1 (662)	1.02 (178)	1.04 (232)	1.06 (64)	1.08 (42)	1.10 (52)	1.12 (30)
Default ($C_0 - C_{12}$)	**0.43**	**0.00**	0.33	**0.21**	0.22	0.26	0.36	0.45	**0.16**	**0.75**	**0.90**	3.31	2.14	**1.43**
$C_0 - C_{11}$	0.43	0.00	0.33	0.28	0.34	0.20	**0.18**	0.36	0.33	0.99	0.90	**2.65**	2.14	1.43
$C_0 - C_{10}$	0.44	0.00	**0.29**	0.42	0.22	0.26	0.18	**0.27**	0.33	1.24	0.90	3.31	2.14	1.43
$C_0 - C_9$	0.51	0.00	0.38	0.35	0.22	0.33	0.54	0.27	0.33	1.24	1.35	3.97	2.86	1.43
$C_0 - C_8$	0.55	0.00	0.48	0.28	0.34	0.20	0.54	0.45	0.33	1.24	1.35	3.97	2.86	1.43
$C_0 - C_7$	0.55	1.16	0.52	0.28	0.34	0.20	0.90	0.31	0.33	1.37	1.35	3.97	1.43	1.43
$C_0 - C_6$	0.53	0.00	0.52	0.35	0.34	**0.13**	0.90	0.31	0.16	1.49	0.90	3.97	1.43	1.43
$C_0 - C_5$	0.59	0.00	0.52	0.28	0.22	0.20	0.90	0.58	0.49	1.37	0.90	3.97	2.14	1.43
$C_0 - C_4$	0.58	0.00	0.57	0.28	0.22	0.33	0.72	0.45	0.49	1.24	1.35	4.64	**0.71**	2.86
$C_0 - C_3$	0.57	0.00	0.57	0.49	**0.11**	0.33	0.90	0.40	0.33	0.99	1.35	3.97	1.43	2.86
$C_0 - C_2$	0.98	0.00	0.95	0.56	0.56	0.79	0.90	0.85	0.82	1.49	1.80	6.62	2.86	2.86

Table 7.9: Performance for adults' test set CAMts on models trained on adults' speech data set CAMtr for various truncations of base MFCC features along with their VTLN warp factor-wise breakup.

MFCC Base Features	% WER										
		VTLN Warp Factor Values									
	All (314)	0.92 (8)	0.94 (14)	0.96 (39)	0.98 (79)	1.00 (70)	1.02 (13)	1.04 (46)	1.06 (8)	1.08 (33)	1.12 (4)
Default $(C_0 - C_{12})$	9.92	6.31	8.87	13.76	8.12	9.82	13.71	**13.93**	4.93	**5.01**	9.76
$C_0 - C_{11}$	**9.76**	6.31	**7.39**	13.30	**7.89**	9.42	13.31	14.55	4.93	5.18	9.76
$C_0 - C_{10}$	10.28	**4.50**	8.37	13.15	8.82	10.06	**11.69**	14.68	10.56	5.87	9.76
$C_0 - C_9$	9.92	5.41	7.88	**12.54**	9.13	**9.19**	**11.69**	14.55	**4.23**	6.22	9.76
$C_0 - C_8$	10.60	5.41	9.36	13.46	9.82	9.50	12.90	15.68	5.63	6.56	**4.88**
$C_0 - C_7$	11.02	5.41	10.84	13.91	10.21	10.30	12.10	15.81	6.34	6.56	7.32
$C_0 - C_6$	11.86	5.41	9.85	15.29	10.90	11.66	12.10	17.31	6.34	6.56	7.32
$C_0 - C_5$	12.95	7.21	12.81	16.97	11.60	12.38	13.71	18.70	**4.23**	7.94	9.76
$C_0 - C_4$	15.08	7.21	15.76	18.96	13.77	14.54	14.11	22.08	9.86	8.64	7.32
$C_0 - C_3$	21.52	13.51	24.14	27.68	20.42	20.69	16.94	28.23	13.38	14.51	17.07
$C_0 - C_2$	29.15	24.32	33.99	37.77	27.38	27.16	25.00	37.39	16.90	21.42	14.63

In order to automate the procedure for determining the appropriate MFCC feature truncation for a test signal on ASR models, it is required to first categorize the input speech as of an adult or of a child based on its degree of gross acoustic mismatch with respect to the recognition models. With respect to adults' speech trained models, the children's speech spectra may require the scaling of frequency axis by a factor as low as 0.88 for normalizing the differences in their formant frequencies which would be unlikely in case of adults' speech spectra. Therefore, the log likelihoods of the default MFCC features of the test signal and of the features corresponding to VTLN warp factor of 0.88 with respect to the adults' speech trained 39-D baseline models, already computed during VTLN warp factor estimation, are compared. The input test speech is categorized as child's speech if its likelihood is more for features corresponding to VTLN warp factor of 0.88 than that corresponding to the default

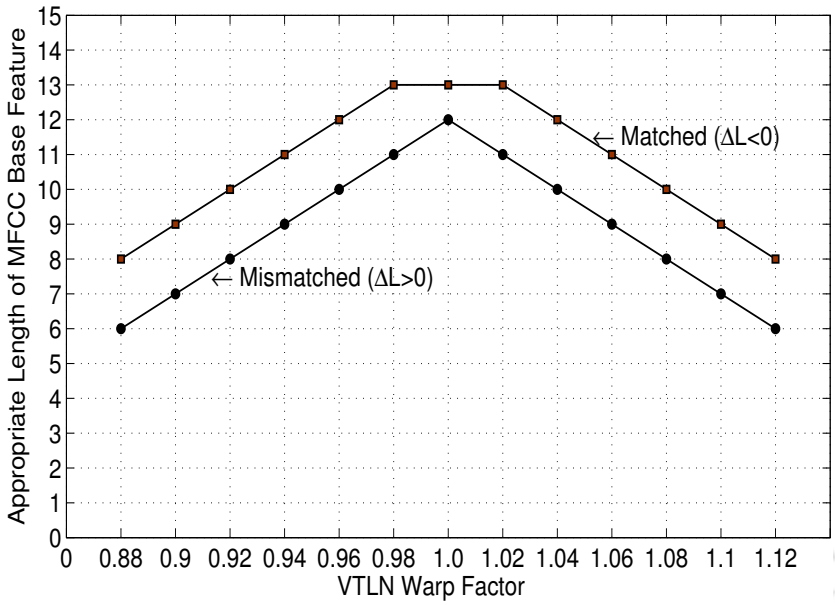


Figure 7.2: Graph showing the proposed relation between the length of base MFCC features and the VTLN warp factor for both matched and mismatched test speech signals.

MFCC features or else as adult's speech. The flow-diagram of the algorithm proposed for ASR on adults' speech trained models is shown in Figure 7.3. Once the input test speech is categorized as adult's or child's speech, the appropriate length of base MFCC features for its recognition is chosen corresponding to its VTLN warp factor estimated with respect to the 39-D baseline models from the piece-wise linear relation given in Figure 7.2.

The recognition performances for both children and adults test sets using the proposed algorithm on the adults' speech trained models on both digit and continuous speech recognition tasks are given in Table 7.10. It is noted that using the proposed algorithm for adaptive truncation of MFCC features relative improvements of 38% and 36% are obtained over baseline in children's ASR performances on adults' speech trained models on the connected digit recognition and the continuous speech recognition tasks. On comparing the improvements obtained in the ASR performances by fixed truncation of MFCC features for all test signals and those using the proposed adaptive MFCC feature truncation algorithm for recognition, it is to note that comparable performances are obtained in both cases on the continuous speech recognition task. The lesser improvement with the proposed adaptive MFCC feature truncation algorithm in comparison to that obtained with fixed truncation of MFCC features for all test signals on connected digit recognition task is attributed to the proposed relation between the VTLN warp factor and the base MFCC feature length. Since, the proposed relation is favored for

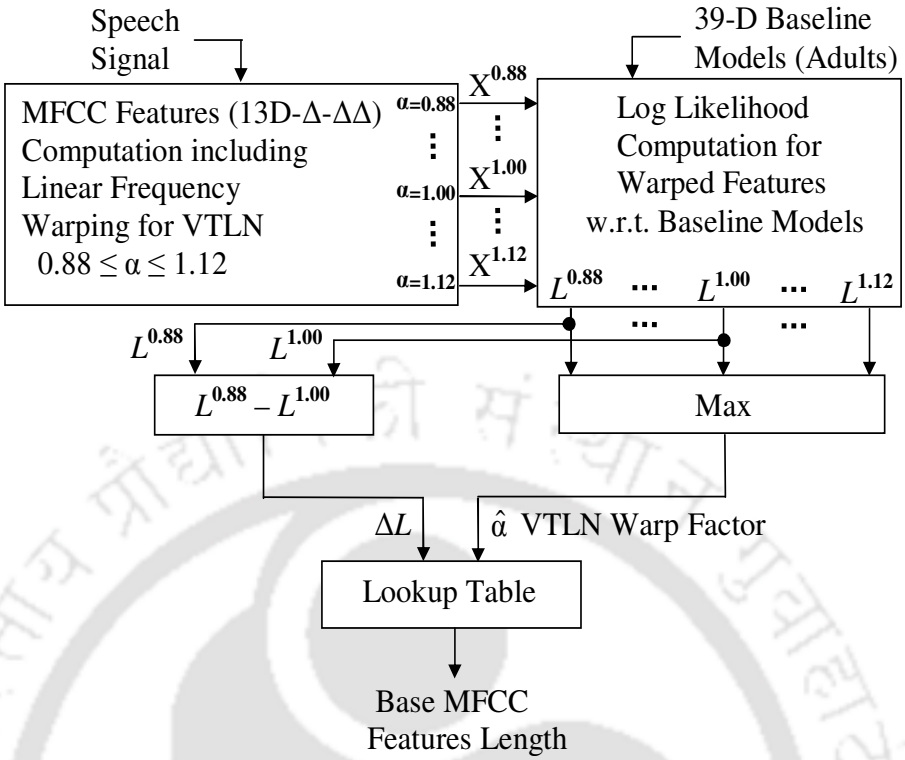


Figure 7.3: Flow diagram of the proposed algorithm to determine the appropriate length of base MFCC features for recognizing a test speech signal on adults' speech trained models. Here the 'Lookup Table' refers to the proposed relation between the length of base MFCC features and the VTLN warp factor shown graphically in Figure 7.2.

the continuous speech recognition task, the improvements obtained on connected digit recognition task using that relation are limited where the children's test speech actually show the need for greater degree of truncation in comparison to that required on the continuous speech recognition task. On recognizing the adults' test speech data, slight insignificant reduction in the ASR performance is observed as compared to its baseline on the adults' speech trained models. Thus, the proposed algorithm is highly efficient in reducing the acoustic mismatch between the children's and the adults' speech data for children's ASR on adults' speech trained models, without using any prior knowledge about the speaker of the test utterance with an additional advantage of reduced MFCC feature dimensions.

7.4.3 Proposed Algorithm for Children's Speech Trained ASR Models

In literature, children's matched ASR performance has been reported to be poorer than the adults' matched ASR performance [60]. This is attributed to the higher inter- and intra-speaker acoustic variability in case of children in comparison to those in case of adults [27, 30]. Therefore, it would be interesting to explore the efficacy of the proposed algorithm in addressing the acoustic mismatch

Table 7.10: Performances for children's and adults' test sets on adults' speech trained models using default MFCC features (referred to as 'Baseline') and MFCC features derived using the proposed algorithm (referred to as 'Proposed') on both connected digit recognition and continuous speech recognition tasks.

Recognition Task	% WER			
	Children's Mismatched ASR		Adults' Matched ASR	
	Baseline	Proposed	Baseline	Proposed
Connected Digit	11.37	7.09	0.43	0.53
Continuous Speech	56.34	36.21	9.92	10.28

for children's matched ASR. With respect to the children's speech trained models, the adults' speech spectra might need expansion by a frequency warp factor of as high as 1.12 which would very exceptionally be required in case of children. So, on children's speech trained models for classifying the test speech signals as of an adult or a child, the log likelihoods of the default MFCC features of the test signal and of the features corresponding to VTLN warp factor of 1.12 with respect to the children's speech trained 39-D baseline models are compared. The input test speech is categorized as adult's speech if the likelihood for the features corresponding to VTLN warp factor of 1.12 is more than that of the default features or else as child's speech. The flow-diagram of the algorithm proposed for ASR on children's speech trained models is shown in Figure 7.4 which is identical to the one proposed for adults' speech trained models shown in Figure 7.3 except for the rule employed to classify the test speech being adult's speech or child's speech.

The recognition performances for both children's and adults' test sets using the proposed algorithm on the children's speech trained models on both digit and continuous speech recognition tasks are given in Table 7.11. It is noted that consistent significant improvements are obtained in the ASR performances for both mismatched adults' test speech and matched children's test speech on both digit and continuous speech recognition tasks. Relative improvements of 49% and 31% are obtained over baseline in children's matched ASR performances on the connected digit recognition and continuous speech recognition tasks, respectively. For adults' speech recognition on children's speech trained models, relative improvements of 35% and 10% are obtained over baseline on the connected digit recognition and continuous speech recognition tasks, respectively. However, it is to note that larger

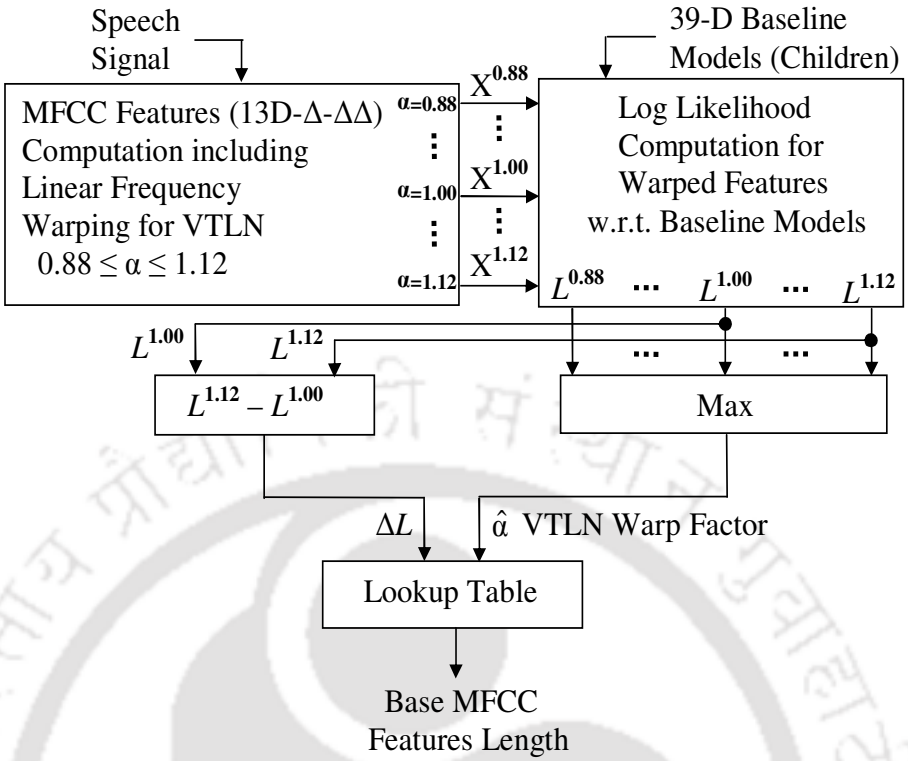


Figure 7.4: Flow diagram of the proposed algorithm to determine the appropriate length of base MFCC features for recognizing a test speech signal on children's speech trained models. Here the 'Lookup Table' refers to the proposed relation between the length of base MFCC features and the VTLN warp factor shown graphically in Figure 7.2.

improvements are obtained for children's speech than for adults' speech on children's speech trained models. Also, for adults' mismatched ASR the improvements are noted to be comparatively less than that for children's mismatched ASR. These are attributed to the poor children's speech trained models having higher variances of the observation densities of phone models due to their higher inter-speaker variability than in case of adults [27,30]. This means that the class-dependent Gaussian densities have more spread and as a result the acoustic classes become less separable in the acoustic feature space in case of children's speech trained models than for adults' speech trained models. This statement can be further substantiated by comparing the average Bhattacharyya distance (BD) [138] between phone classes for adults's and children's speech trained phone models.

Given two Gaussian distributions $\mathcal{N}(\mu_i, \Sigma_i)$ and $\mathcal{N}(\mu_j, \Sigma_j)$ representing phone 'i' and phone 'j', respectively, the BD between these distributions is computed as:

$$BD(i,j) = \frac{1}{8}(\mu_i - \mu_j)^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1}(\mu_i - \mu_j) + \frac{1}{2} log \frac{\left|\frac{\Sigma_i + \Sigma_j}{2}\right|}{\sqrt{|\Sigma_i||\Sigma_j|}} \quad (7.5)$$

Table 7.11: Performances for children's and adults' test sets on children's speech trained models using default MFCC features and MFCC features derived using the proposed algorithm on both connected digit recognition and continuous speech recognition tasks.

Recognition Task	% WER			
	Children's Matched ASR		Adults' Mismatched ASR	
	Baseline	Proposed	Baseline	Proposed
Connected Digit	1.01	0.52	13.28	8.70
Continuous Speech	12.41	8.62	68.36	61.43

Given a set of M Gaussian densities representing M phone classes, the average BD is determined as follows:

$$AvgBD = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} BD(i,j) \tag{7.6}$$

The average BD, $AvgBD$, can be considered as a statistical measure of how scattered the M phones are in the acoustic space. High values of $AvgBD$ indicate that phone distributions are well scattered in the acoustic space and thus phones should be more easily discriminated, while low values of $AvgBD$ can be interpreted as a higher superposition of phone distributions and thus the phone discrimination task would become harder.

To compare the inter-speaker variability of adults' and children's speech, the average BD is measured for both adults' and children's speech trained phone models. The phone models are built using a 3-state left-to-right topology with a single Gaussian density per state. Each speech frame is parameterized by a 39-D observation vector composed of 13 MFCCs ($C_0 - C_{12}$) plus their first and second order temporal derivatives. Cepstral mean subtraction is performed on static features on an utterance-by-utterance basis. For children, two sets of phone models are trained each using CHtr and PFtr data sets while for adults, the two sets of phone models are trained using ADtr and CAMtr data sets. In computing the average BD, only the Gaussian densities associated to the central states of the phone models are considered. This is done based on our assumption that the Gaussian density associated to the central state of a model better reflects the acoustic characteristics of the modeled phone than Gaussian densities associated to the initial and final states.

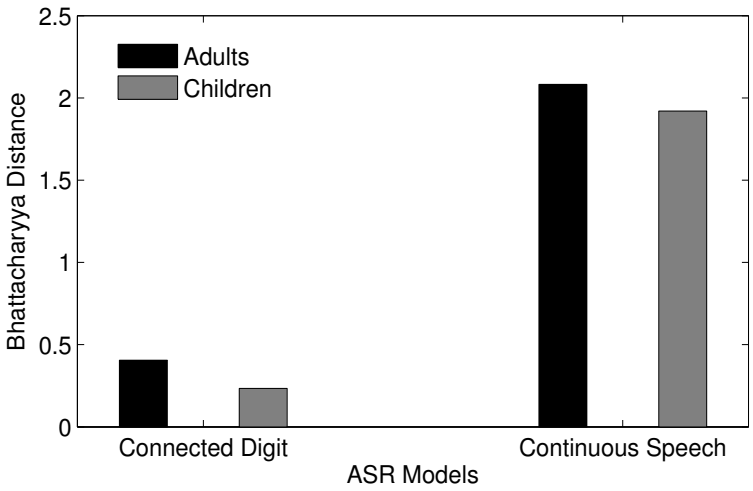


Figure 7.5: Bar graph showing average Bhattacharyya distance across vowel sounds for models trained on adults' and children's speech for connected digit and continuous speech recognition tasks.

The average BD for the adults' and the children's speech trained phone models is shown in Figure 7.5. Seven different vowels are considered in computing BD measures in all cases. It can be noted that the average BD among vowel distributions is greater for adults' speech trained models than for children's speech trained models showing that vowel distributions are more overlapped in the acoustic spaces of children's speech leading to comparatively poor classification performance of children's speech trained models. This is attributed to the effect of greater inter-speaker acoustic variability in children's speech than in adults' speech [27].

7.5 Combining Proposed Algorithm with VTLN and CMLLR

From the previous sections, it is noted that MFCC feature truncation also addresses the acoustic mismatch and thus, our proposed algorithm for adaptive MFCC feature truncation gives improvements in the acoustically mismatched ASR performance. So, it would be interesting to explore whether the improvement obtained by the proposed algorithm is additional to that obtained by the existing speaker normalization and model adaptation techniques viz., VTLN and CMLLR [30, 69] or not. The VTLN and CMLLR are performed on the test data using HTK as described in Chapter 2.

The recognition performances for the children's test set PFts on the adults' speech trained models and the children's speech trained models using the default MFCC features and the features derived using the proposed adaptive MFCC feature truncation algorithm both with and without VTLN are given in Table 7.12. It is noted that further 15% relative improvement is obtained in the children's ASR

Table 7.12: Performance for PFts test set using the default MFCC features referred to as 'Default' and MFCC features derived using the proposed algorithm referred to as 'Proposed' both with and without VTLN and CMLLR under both matched and mismatched conditions.

Condition	Children's Mismatched ASR			Children's Matched ASR		
	% WER		% Relative Gain with CMLLR	% WER		% Relative Gain with CMLLR
	Baseline	with CMLLR		Baseline	with CMLLR	
Default	56.34	38.25	32	12.41	10.22	18
Default + VTLN	26.78	18.63	30	9.06	7.99	12
Proposed	36.21	23.07	36	8.62	7.54	13
Proposed + VTLN	22.72	**16.16**	29	7.70	**6.75**	12

performance by doing VTLN using MFCC features derived using the proposed adaptive truncation algorithm over the performance obtained with VTLN using the default features on both the adults' speech trained and the children's speech trained models. However, it is to note that the relative improvement with VTLN is reduced by 15% when the MFCC features derived using the proposed algorithm are used in comparison to that obtained in case of the default MFCC features. This is attributed to reduction of the acoustic mismatch due to the VTL differences with truncation of higher order coefficients in MFCC features as argued earlier.

The recognition performances for the children's test set PFts on the adults' speech trained and the children's speech trained models using the default MFCC features and the features derived using the proposed adaptive MFCC feature truncation algorithm both with and without CMLLR are also given in Table 7.12. It is noted that large relative gains of 40% and 26% are obtained in the children's ASR performance on doing CMLLR using MFCC features derived using the proposed algorithm over the performance obtained by doing CMLLR using the default MFCC features on the adults' speech trained and the children's speech trained models, respectively. On further noting the recognition performances obtained by combined VTLN and CMLLR, it is found that relative improvements of 13% and 16% are obtained in the children's ASR performance when MFCC features derived using the proposed algorithm are used over that obtained in case of the default MFCC features on the adults' speech trained and the children's speech trained models, respectively.

Thus, the improvement obtained using the proposed MFCC feature truncation algorithm is additive to those obtained with VTLN and CMLLR. Like CMLLR, VTLN can also be implemented through application of linear transformations to MFCC features [139, 140]. So, the improvement obtained with the proposed adaptive MFCC feature truncation algorithm being additive to those obtained with VTLN and CMLLR is attributed to the former not being constrained to linear transformations unlike the latter.

It is to note that by reducing the feature dimensions for both the training and the test speech data at the same time, the speaker adaptive training is implicitly incorporated in the proposed algorithm and thus, does not require to explicitly adapt and re-train the models accordingly. In addition to this, with reduction in MFCC feature dimensionality, the proposed technique also leads to reduction of computational complexity involved in decoding as well as in doing subsequent VTLN/CMLLR adaptation compared to that in case of the default MFCC features.

7.6 Summary

Motivated by the effect of pitch on higher order coefficients of MFCC ($C_0 - C_{12}$) feature vector in case of high pitch signals noted in Section 3.2.1 and Section 4.3, in this chapter, MFCC feature truncation is explored for children's speech recognition on the adults' speech trained models. Following the observations, the contribution of each of the coefficients of the 39-D MFCC feature vector in the children's ASR performance on the adults' speech trained models is explored individually. The work done in this chapter and the observations made in this study are summarized below.

- Significantly large improvement is obtained in the children's ASR performance with increased truncation of base MFCC features in comparison to the default length of 13.
- On exploring the role of MFCC feature truncation in the pitch mismatch reduction, it is noted that the Mahalanobis distance of the high pitch signals with respect to low pitch speech trained recognition models is significantly reduced on truncating the higher order coefficients of MFCC ($C_0 - C_{12}$) features due to reduction in their pitch mismatch.
- With increased truncation, the VTL differences are also significantly reduced due to additional spectral smoothing.

- An automatic algorithm is proposed for utterance-specific truncation of MFCC features of test signals for reducing their pitch mismatch with respect to the ASR models without any prior knowledge about the speaker of the test utterance.
- Significantly large improvements are obtained in the children's speech recognition performances on the adults' speech trained models using the proposed algorithm similar to that corresponding to the best performance obtained by using fixed feature truncation for all test signals.
- Using the proposed algorithm, significant improvements are found in the children's matched and adults' mismatched ASR performances as well with slight degradation in the adults' matched ASR performance.
- The improvements obtained in the ASR performances with the proposed algorithm are also found to be additive to those obtained with the existing speaker normalization and model adaptation techniques viz., VTLN and CMLLR with an additional advantage of reduction in the MFCC feature dimensionality.

8

Summary and Future Work

Contents

8.1 Summary of the Work 138
8.2 Contributions of the Work 140
8.3 Scope for the Future Work 141

8.1 Summary of the Work

The main objective of this thesis is to characterize the pitch mismatch between the adults' and the children's speech and to develop techniques for its reduction for improving children's ASR performance on adults' speech trained models.

In the beginning of the work, the effect of differences in each of the acoustic correlates of speech viz., the pitch, the speaking rate, the formant frequencies and the glottal flow parameters (open quotient, return quotient and speed quotient) across speech signals on MFCC features and on the HMM-based ASR models is explored. The relative significance of each of these acoustic sources of mismatch for children's ASR on adults' speech trained models is determined by explicitly normalizing the differences in each of these acoustic correlates in children's test speech signals with respect to the adults' speech trained models. It is found that, apart from the formant frequencies, the pitch is the other major source of acoustic mismatch that significantly affects the children's speech recognition performance. The variances of the higher order coefficients of the 13-D default base MFCC features ($C_0 - C_{12}$) are found to significantly increase with increase in the pitch of the signals. Significant improvement is obtained in the children's ASR performance with explicit pitch normalization of children's speech which is also shown to be additive to those obtained with the existing model adaptation techniques viz., MLLR and CMLLR.

Motivated by the observed increase in variances of the higher order coefficients of 13-D base MFCC features with increase in the pitch of the signals, the cause and the nature of the effect of pitch on MFCC features is then explored in detail. It is shown with the help of the real and the synthetic examples that some pitch-dependent distortions appear in the Mel spectral envelope for high pitch signals due to insufficient smoothing of the pitch harmonics in the speech spectrum by the non-uniform Mel filterbank. However, on account of the constant-Q type Mel filterbank used in MFCC feature computation, there in no harmonicity in the pitch-dependent distortions appearing in the Mel spectrum. As a result, no harmonicity appears in the Mel cepstrum corresponding to those pitch-dependent distortions. Thus, the effect of these pitch-dependent distortions appears on all Mel cepstral coefficients and can not be eliminated completely by cepstral truncation. Further, it is shown that the effect of these pitch-dependent distortions on the default 13-D base MFCC features is comparatively more on the higher order coefficients than on the lower order coefficients.

On observing the significant degradation in the children's ASR performance due to pitch differences

using MFCC features, we explored the pitch-robustness of other salient features viz., PLPCC and PMVDR for children's ASR in comparison to MFCC features. It is found that PLPCC features are affected by the pitch variations across speech signals to a lesser extent compared to MFCC features. However, after explicit pitch normalization of children's speech better performance is obtained with MFCC features compared to PLPCC features for children's ASR on adults' speech trained models. On the other hand, with suitable optimization of model order PMVDR features are found to be more pitch-robust than default MFCC features. However, the children's ASR performance obtained with PMVDR features after optimization of its model order for children's speech is found to be comparable to that obtained with MFCC features after explicit pitch normalization of children's speech.

ML-based explicit pitch normalization approach is computationally expensive and its performance is subject to the accuracy of the pitch marks. Also, PMVDR feature computation requires *twice* the number of operations compared to that for MFCC feature computation. Therefore, further, suitable modifications are explored for MFCC feature computation for improving children's speech recognition on adults' speech trained models without explicit pitch normalization of children's speech.

First, an algorithm is proposed for normalizing the pitch differences across speech signals during MFCC feature extraction itself. The proposed pitch normalization algorithm increases the bandwidths of selected filters in the Mel filterbank so as to effect proper smoothing of the pitch harmonics in the speech spectrum. The filterbank structure is modified during MFCC feature extraction for each children's test speech signal according to the average pitch of the test signal using the proposed pitch normalization algorithm. Significant relative improvements of 16% (on connected digit recognition task) and 9% (on continuous speech recognition task) are obtained in the children's ASR performance using the proposed filterbank based method for pitch normalization comparable to those obtained with explicit pitch normalization. The proposed algorithm for utterance-specific modification of the bandwidths of only selected filters in the filterbank is noted to give larger improvement than those obtained in earlier works either by explicitly modifying the bandwidths of all filters or by reducing the number of filters in the filterbank. The improvements with the proposed algorithm are also found to be additive to those obtained with the existing speaker normalization and model adaptation techniques viz. VTLN and CMLLR for children's ASR.

Later, on observing the effect of pitch on higher order coefficients of MFCC features in case of high pitch signals, further truncation of 13-D MFCC ($C_0 - C_{12}$) features is explored for children's speech

recognition on the adults' speech trained models. It is found that significantly large improvement is obtained in the children's ASR performance with increased truncation of MFCC features in comparison to the default base feature length of 13. On further exploring the role of MFCC feature truncation in the pitch mismatch reduction, it is noted that the Mahalanobis distance of the high pitch signals with respect to low pitch speech trained recognition models is significantly reduced on truncating the higher order coefficients of MFCC features due to reduction in their pitch mismatch. Also, it is noted that with increased truncation, the vocal tract length differences are also significantly reduced due to the additional spectral smoothing resulting on account of cepstral truncation.

Based on these observations, an automatic algorithm is proposed for utterance-specific truncation of 13-D base MFCC features of test signals for reducing their pitch mismatch with respect to the ASR models without any prior knowledge about the speaker of the test utterance. Significantly large relative improvements of 38% (on connected digit recognition task) and 36% (on continuous speech recognition task) are obtained in the children's speech recognition performances on the adults' speech trained models using the proposed algorithm. Significant relative improvements are found in the children's matched ASR performance (49% and 31% on connected digit recognition and continuous speech recognition tasks, respectively) and adults' mismatched ASR performance (35% and 10% on connected digit recognition and continuous speech recognition tasks, respectively) as well with slight degradation in the adults' matched ASR performance. Interestingly, the improvements obtained in the ASR performances with the proposed algorithm are also found to be additive to those obtained with the existing speaker normalization and model adaptation techniques viz., VTLN and CMLLR with an additional advantage of reduction in the MFCC feature dimensionality.

8.2 Contributions of the Work

In this work, an effort is made to address the pitch mismatch in context of children's ASR. The salient contributions made in this thesis are:

- The relative significance of various acoustic sources of mismatch (pitch, speaking rate, formant frequencies, glottal flow parameters) has been determined for children's ASR on adults' speech trained models. It has been shown that, from the perspective of ASR, the pitch is the second most important source of acoustic mismatch between adults' and children's speech after formant frequencies.

- The effect of pitch has been characterized on the most commonly used MFCC features. It is shown that due to insufficient smoothing of the pitch harmonics by the non-uniform Mel filterbank, some non-harmonic pitch-dependent distortions appear in the smoothed Mel spectrum. Also, it is shown that the effect of these pitch-dependent distortions which is comparatively more on the higher order coefficients than on the lower order coefficients of the 13-D base MFCC features can not be eliminated completely by cepstral truncation.

- The pitch-robustness of PLPCC and PMVDR features has been explored in context of their efficacy for children's ASR on adults' speech trained models in comparison to MFCC features. It is shown that though with suitable optimization of model order PMVDR features are more pitch-robust than MFCC and PLPCC features, they give children's ASR performance comparable to that obtained with MFCC features after explicit pitch normalization of children's speech.

- An utterance-specific pitch adaptive Mel filterbank modification algorithm has been proposed for pitch normalization of children's speech for their recognition on adults' speech trained models using MFCC features. The improvement obtained in the children's ASR performance using the proposed Mel filterbank adaptation algorithm is comparable to that obtained with explicit pitch normalization approach.

- The role of MFCC feature truncation in pitch mismatch reduction has been identified and demonstrated experimentally. It is noted that significant improvement is obtained in the ASR performance for high pitch signals on low pitch speech trained recognition models on truncating the higher order coefficients of MFCC features.

- An utterance-specific MFCC feature truncation algorithm has been proposed for pitch mismatch reduction between training and test speech data without prior knowledge about the speaker of the test utterance for ASR. A large improvement is obtained in the ASR performance under acoustically mismatched condition with much reduced computational complexity using the proposed truncation algorithm.

8.3 Scope for the Future Work

The proposed algorithms in this work are shown to provide significant improvements for children's (high pitch) speech recognition on adults' (low pitch) speech trained acoustic models for neutral speech

case. It is already known from literature that in comparison to neutral speech pitch is significantly different for stressed/emotional speech. For instance, it is reported in [141] that the mean pitch values of angry speech are around 100-150 Hz more than those of neutral speech. So, the degradation in the ASR performance for stressed/emotional speech on models trained on neutral speech are partly attributed to the differences in their pitch values. The proposed filterbank based pitch normalization algorithm addresses the pitch mismatch and thus, can be explored for stressed/emotional speech recognition on ASR models trained on neutral speech.

The proposed MFCC feature truncation based algorithm for pitch mismatch reduction gives large improvement in children's ASR on adults' speech trained models. However, increase in cepstral truncation might cause the loss of the relevant spectral information. To avoid this, the heteroscedastic linear discriminant analysis (HLDA)-based transformations can be explored for optimal selection of base MFCC features length. Interestingly, a study has been reported which explores combination of CMLLR and HLDA for speaker adaptation for adults' ASR [142]. Similar study, thus, can be explored for addressing the acoustic mismatch for children's ASR which is expected to further enhance the performance of the children's speech recognition on adults' speech trained models.

In this thesis, all ASR evaluations have been done using narrowband speech data. From wideband to narrowband speech, greater loss of spectral information occurs in the children's speech spectra in comparison to adults' speech spectra [65, 66]. So, the studies reported in this thesis may be reviewed for children's speech recognition on ASR systems using wideband speech data. The work in this thesis is to address the pitch mismatch whose effect has been noted predominantly on the lower frequency region of the speech spectrum. So, most of the observations made in this work would remain same with increase in the signal bandwidth. But, it would be interesting to explore the effect of signal bandwidth particularly on the correlation between the pitch mismatch and the MFCC features length used for developing the adaptive MFCC feature truncation algorithm proposed in this thesis.

A

Mel Frequency Cepstral Coefficients

Contents

A.1 Mel Frequency Cepstral Coefficient (MFCC) Computation 144

A.1 Mel Frequency Cepstral Coefficient (MFCC) Computation

In a typical Mel frequency cepstral coefficient (MFCC) feature extraction process [90], the first step is to pre-emphasize and windowing the speech signal to divide it into short-time frames. After windowing, discrete Fourier transform (DFT) is used to find the magnitude spectrum of each frame. Here we perform filter bank processing to the magnitude spectrum, which uses Mel scale. Discrete cosine transformation (DCT) is then applied to the log of the Mel-warped magnitude spectrum to compute MFCC coefficients. The detailed description of various steps involved in MFCC feature extraction is explained below.

(i) **Pre-emphasis:** The speech signal has an overall spectral slope of -6dB per octave due to the combined effect of glottal pulse roll-off (-12dB per octave [143]) and the lip radiation (+6dB per octave). This slope is compensated by performing pre-emphasis on the speech signal using a pre-emphasis filter. The most commonly used pre-emphasis filter is given by the following transfer function:

$$H(z) = 1 - az^{-1} \tag{A.1}$$

where, the value of a controls the slope of the filter and is usually between 0.9 to 1.0.

(ii) **Frame Blocking and Windowing:** The speech signal is a slowly time-varying or quasi-stationary signal. Therefore, speech analysis is always carried out on short-time speech segments across which the speech signal is assumed to be stationary. Short-time spectral measurements are typically carried out using an analysis window of length equal to 20-25 ms. Usually, the frames are set to overlap so that their centers lie only 10 ms apart. This enables the temporal characteristics of individual speech sounds to be tracked. Generally, Hanning or Hamming window [143], is used for analysis so that the values near the edges become zero. This is done to enhance the harmonics, smooth the edges and to reduce the discontinuities at the edges while taking the DFT on the signal.

(iii) **DFT Spectrum:** Each windowed speech frame is converted into its frequency domain representation by applying DFT as given in Eqn. A.2 [144]. This results in a short-term magnitude spectrum of a speech frame. Here phase information is discarded because it does not carry useful information from the hearing perspective.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nk}{N}} \quad \text{for } 0 \leq k \leq N-1 \tag{A.2}$$

where, N is the number of points used to compute the DFT.

(iv) **Mel Filterbank Processing:** A Mel-warped spectrum is computed by passing the Fourier transformed signal through a set of band-pass filters known as Mel filterbank. A Mel is a unit of measure based on the human ear's perceived frequency. The Mel scale is, therefore, a mapping between the physical frequency scale (Hz) and the perceived frequency scale (Mel). The Mel scale has approximately a linear frequency spacing below 1 kHz and a logarithmic spacing above 1 kHz [145]. The approximation of Mel from physical frequency can be expressed as [146]:

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{A.3}$$

where, f denotes the physical frequency in Hz, and f_{mel} denotes the perceived frequency.

Filter banks can be implemented in both time domain and frequency domain. For MFCC computation, filter banks are generally implemented in frequency domain. The center frequencies of the filters are normally evenly spaced on the frequency axis. However, in order to mimic the human ears' perception, the warped axis according to the non-linear function given in Eqn. A.3, is implemented. The bandwidths of the filters are decided based on the critical bandwidth phenomena [2] noted in the psychoacoustic studies for human auditory perception such that each pair of consecutive filters have 50% overlapping. The most commonly used filter shaper is triangular, and in some cases the Hanning filter can be found [143].

The Mel-warped spectrum of the Fourier transformed magnitude spectrum $X(k)$ is computed by multiplying the magnitude spectrum by each of the of the triangular Mel weighting filters as:

$$s(m) = \sum_{k=0}^{N-1} \left[|X(k)|^2 H_m(k) \right] \quad \text{for } 0 \leq m \leq M-1 \tag{A.4}$$

where, M is total number of triangular Mel weighting filters [147].

(v) **Discrete Cosine Transform (DCT):** Since the vocal tract is smooth, the energy levels in adjacent bands tend to be correlated. A set of de-correlated cepstral coefficients are obtained by applying DCT to the logarithm of the transformed Mel frequency coefficients. Since DCT

gathers most of the information in the signal to its lower order coefficients, by discarding the higher order coefficients significant reduction in computational cost and robustness of systems can be achieved [143]. Typically, first 13 coefficients are chosen for speech recognition. Finally, MFCC is calculated as [143]:

$$c(n) = \sum_{m=0}^{M-1} \log_{10}\left(s(m)\right) \cos\left(\frac{\pi n(m-0.5)}{M}\right) \; ; \qquad n = 0, 1, 2, ..., C-1 \tag{A.5}$$

where, $c(n)$ are the cepstral coefficients and C is the number of MFCCs. The logarithm of the energy for the frame (i.e., the zeroth order coefficient) is added to Mel frequency cepstral coefficients because different phonemes may have different energy.

(vi) **Dynamic MFCC Features:** The cepstral coefficients described so far are referred to as the static features, since they capture only the average frequency distribution for a given frame. The extra information about the temporal dynamics of the signal is obtained by computing first and second derivatives of cepstral coefficients [148–150]. The first order derivative is called delta coefficients and the second order derivative is called delta-delta coefficients. Delta coefficients tells about the speech rate and the delta-delta coefficients gives an information similar to acceleration of speech.

The commonly used definition for computing dynamic parameter is [148]:

$$\Delta c_m(n) = \frac{\sum_{i=-T}^{T} k_i c_m(n+i)}{\sum_{i=-T}^{T} |i|} \tag{A.6}$$

where, $c_m(n)$ denotes the m^{th} feature for the n^{th} time frame, k_i is the i^{th} weight and T is the number of successive frames used for computation. Generally, T is taken as 2.

B

Hidden Markov Models

Contents

B.1 Hidden Markov Model (HMM) 148

B.1 Hidden Markov Model (HMM)

Hidden Markov model (HMM) is the mainstream of acoustic modeling in almost all practical large vocabulary ASR systems [151]. In HMM based acoustic modeling, the acoustic model consists of an HMM for each of the basic modeling unit, e.g., a word for a small vocabulary of training speech data. For large vocabulary ASR, usually an HMM is constructed for each phone. The HMM of a word is then constructed by concatenating corresponding phone-specific HMMs. We can further concatenate HMMs of words to construct the HMM of the whole string that contains multiple words.

Given the observation vector $\mathbf{o}_t$ at time t, the parameter set that defines a continuous density HMM, λ, having S discrete states where each state is modeled by a M-mixtures Gaussian density function, comprises of:

(i) $\mathbf{A} = \{a_{ij}\ ;\quad 1 \leq i, j \leq S\}$, transition probabilities between two states

(ii) $\mathbf{B} = \{b_j(\mathbf{o}_t)\ ;\quad 1 \leq j \leq S\}$, observation probabilities of states

(iii) $\pi = \{\pi_i\ ;\quad 1 \leq i \leq S\}$, initial probabilities of the states

A typical left-to-right HMM topology, represented by two special non-emitting states in a HMM, is used in the ASR systems. The two non-emitting states include an entry state and an exit state which are reached only once. Because they do not generate any observation, none of them has an emitting probability density. The initial probabilities are, thus, simply $\pi_1 = 1$ and $\pi_i = 0$ for $i \neq 1$.

The algorithms used for training (setting parameters of the model from observations) and testing (decoding a test utterance to hypothesize a word string) the acoustic models of the continuous density HMM-based isolated unit ASR systems are described below.

B.1.1 Training

The parameters of the HMMs viz., the state transition probabilities, the Gaussian mixture weights and the means and variances of the Gaussian distributions are learned in a data-driven manner using the Baum-Welch re-estimation algorithm [152]. The algorithm performs the maximum likelihood training of HMM parameters using a variation of the expectation-maximization (EM) algorithm [153]. Given an initial model λ and the observation sequence $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T\}$, each iteration increases the likelihood of the observation sequence i.e., $P(\mathbf{O}|\lambda)$ till it converges to a local maximum. The function $P(\mathbf{O}|\lambda)$ is called the likelihood function.

First, a rough guess of the parameter values $\mathbf{A}$, $\mathbf{B}$ and π of an initial model λ is made, using either a *flat start* training or a segmental K-means algorithm [154]. The segmental K-means algorithm for initialization of models is implemented as:

$$\pi = \{\pi_i\} \tag{B.1}$$

where, $\pi_i = P(i_1 = i)$, Probability of being in state i at $t = 1$.

$$\mathbf{A} = \{a_{ij}\} \tag{B.2}$$

where, $a_{ij} = P(i_{t+1} = j | i_t = i)$, Probability of being in state j at time $t + 1$ given that we were in state i at time t. a_{ij}'s are assumed to be independent of time.

$$\mathbf{B} = \{b_j(\mathbf{o}_t)\} \tag{B.3}$$

where, $b_j(\mathbf{o}_t) = \sum_{m=1}^{M} c_{jm} \mathcal{N}(\mathbf{o}_t; \mu_{jm}, \mathbf{\Sigma}_{jm})$, Output probability of the observation $\mathbf{o}_t$ given that we are in state j. Here,

$$\mathcal{N}(\mathbf{o}_t; \mu_{jm}, \mathbf{\Sigma}_{jm}) = \frac{1}{\sqrt{(2\pi)^D |\mathbf{\Sigma}_{jm}|}} e^{-\frac{1}{2}(\mathbf{o}_t - \mu_{jm})^T \mathbf{\Sigma}_{jm}^{-1} (\mathbf{o}_t - \mu_{jm})} \tag{B.4}$$

is a multivariate Gaussian density with D representing the dimension of the feature vector $\mathbf{o}_t$ and c_{jm}, μ_{jm} and $\mathbf{\Sigma}_{jm}$ are the weight, mean vector, and covariance matrix of the m^{th} Gaussian component of the mixture distribution in state j, respectively.

On the other hand, in *flat start* training, all models are initialized to be identical with means and variances of each state being equal to the global mean and variance of training speech data. In this thesis, the parameters of the HMM-based models are initialized using the *flat start* training approach. Given the observation sequence $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T\}$ and the model $\lambda = (\mathbf{A}, \mathbf{B}, \pi)$, subject to the stochastic constraints

$$\sum_{j=1}^{S} \pi_j = 1 \tag{B.5}$$

$$\sum_{j=1}^{S} a_{ij} = 1 \quad for \;\; 1 \leq i \leq S \tag{B.6}$$

$$\sum_{m=1}^{M} c_{jm} = 1 \quad for \;\; 1 \leq j \leq S \tag{B.7}$$

$$\int_{-\infty}^{\infty} b_j(\mathbf{o}_t) d\mathbf{o}_t = 1 \quad for \;\; 1 \leq j \leq S \tag{B.8}$$

more accurate parameter values of the model are found by using the Baum-Welch re-estimation formulae:

$$\hat{\pi_i} = \frac{\alpha_1(i)\beta_1(i)}{\sum_{j=1}^{S} \alpha_T(j)} = \gamma_1(i) \tag{B.9}$$

$$\hat{a_{ij}} = \frac{\sum_{t=1}^{T-1} \alpha_t(i) a_{ij} b_j(\mathbf{o}_t) \beta_t(j)}{\sum_{t=1}^{T-1} \alpha_t(i)\beta_t(i)} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{B.10}$$

$$\hat{c_{jm}} = \frac{\sum_{t=1}^{T} \gamma_t(j,m)}{\sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_t(j,m)} \tag{B.11}$$

$$\hat{\mu_{jm}} = \frac{\sum_{t=1}^{T} \gamma_t(j,m).\mathbf{o}_t}{\sum_{t=1}^{T} \gamma_t(j,m)} \tag{B.12}$$

$$\hat{\mathbf{\Sigma}_{jm}} = \frac{\sum_{t=1}^{T} \gamma_t(j,m).(\mathbf{o}_t - \mu_{jm})(\mathbf{o}_t - \mu_{jm})^T}{\sum_{t=1}^{T} \gamma_t(j,m)} \tag{B.13}$$

where,

$$\gamma_t(i) = P(i_t = i|\mathbf{O}, \lambda) \tag{B.14}$$

defines the probability of being in state i at time t,

$$\gamma_t(j,m) = \left[\frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{S} \alpha_t(j)\beta_t(j)}\right] \left[\frac{c_{jm}\mathcal{N}(\mathbf{o}_t; \mu_{jm}, \mathbf{\Sigma}_{jm})}{\sum_{m=1}^{M} \mathcal{N}(\mathbf{o}_t; \mu_{jm}, \mathbf{\Sigma}_{jm})}\right] \tag{B.15}$$

defines the probability of being in state j at time t with the m^{th} mixture component accounting for $\mathbf{o}_t$, and

$$\xi_t(i,j) = P(i_t = i, i_{t+1} = j|\mathbf{O}, \lambda) \tag{B.16}$$

defines the probability of being in state i at time t and making a transition to state j at time $t+1$. Hence,

$$\sum_{t=1}^{T-1} \gamma_t(i) = \text{Expected number of transitions from state i} \tag{B.17}$$

$$\sum_{t=1}^{T-1} \xi_t(i,j) = \text{Expected number of transitions from state i to state j} \tag{B.18}$$

Therefore, the re-estimation formula for π_i is the probability of being in state i at time $t = 1$. The formula for a_{ij} is the ratio of expected number of times of making a transition from state i to state j to the expected number of times of making a transition out of state i.

B.1.2 Testing

For an acoustic signal, given the observation sequence $\mathbf{O}$, the speech recognition using HMMs employs Bayes rule is given as:

$$\hat{\lambda} = \arg\max_{\lambda} P(\lambda|\mathbf{O}) = \frac{P(\lambda)P(\mathbf{O}|\lambda)}{P(\mathbf{O})} \tag{B.19}$$

which finds the most likely model λ for the given observation sequence $\mathbf{O}$. Here, $P(\lambda)$ determines the probability of model which is estimated using language models, $P(\mathbf{O}|\lambda)$ is the conditional probability of the occurrence of the observation sequence $\mathbf{O}$ given the model λ and $P(\mathbf{O})$ is the probability of the observation sequence which is independent of the model λ. Therefore, the problem is to find the maximum value of the product of $P(\lambda)$ with $P(\mathbf{O}|\lambda)$. This requires to determine the maximum value of the probability $P(\mathbf{O}|\lambda)$ across all trained models. Since, the observations are generated by states which are hidden, it is required to determine the hidden state sequence that can generate the observation sequence $\mathbf{O}$ given the model λ.

Thus, in decoding, given the model $\lambda = (\mathbf{A}, \mathbf{B}, \pi)$, the problem is to find a state sequence $\mathbf{I} = \{i_1, i_2, ..., i_T\}$ so that the joint probability of the observation sequence $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T\}$ and the state sequence given the model is maximized. For identifying the underlying hidden state sequence $\mathbf{I}$ that maximizes $P(\mathbf{O}, \mathbf{I}|\lambda)$, a dynamic programming algorithm known as the Viterbi algorithm [152] is used. It is an inductive algorithm in which at each instant you keep the state sequence giving the maximum probability for each of the S states as the intermediate state for the desired observation sequence $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T\}$. In this way, you finally have the best path for each of the S states as the last state for the desired observation sequence. Out of these, we select the one which has the highest probability. The step-by-step implementation of the Viterbi algorithm is described below:

(i) **Initialization:** For $1 \leq i \leq S$

$$\delta_1(i) = -\ln(\pi_i) - \ln(b_i(\mathbf{o}_1)) \tag{B.20}$$

$$\psi_1(i) = 0 \tag{B.21}$$

(ii) **Recursive Computation:** For $2 \leq t \leq T$ and $1 \leq j \leq S$

$$\delta_t(j) = \min_{1 \leq i \leq S} [\delta_{t-1}(i) - \ln(a_{ij})] - \ln(b_j(\mathbf{o}_t)) \tag{B.22}$$

$$\psi_t(j) = \arg \min_{1 \le i \le S} [\delta_{t-1}(i) - \ln(a_{ij})] \tag{B.23}$$

(iii) **Termination:**

$$P^* = \min_{1 \le i \le S} [\delta_T(i)] \tag{B.24}$$

$$q_T^* = \arg \min_{1 \le i \le S} [\delta_T(i)] \tag{B.25}$$

(iv) **Tracing back the optimal state sequence:** For $t = T-1, T-2, ..., 1$

$$q_t^* = \psi_{t+1}(q_{t+1}^*) \tag{B.26}$$

Hence, $exp(-P^*)$ gives the required state-optimized probability, and $Q^* = \{q_1^*, q_2^*, ..., q_T^*\}$ is the optimal state sequence.

C

Perceptual Linear Prediction Cepstral Coefficients

Contents

C.1 Perceptual Linear Prediction Cepstral Coefficient (PLPCC) Computation 154

C.1 Perceptual Linear Prediction Cepstral Coefficient (PLPCC) Computation

Perceptual linear prediction (PLP), introduced in [122], uses several concepts from psychophysics of hearing for frequency weighting to determine an estimate of auditory spectrum. The auditory spectrum is then approximated by an autoregressive low-order all pole model. As a result, PLP spectrum does not reflect speaker-dependent details of the spectrum of speech, merging the higher resonance spectral peaks.

The basic steps involved in a typical perceptual linear prediction cepstral coefficient (PLPCC) feature computation process are described below.

(i) **Spectral Analysis:** As speech is quasi-stationary, the speech signal is divided into frames and weighted by an analysis window, which is often a Hamming window.

$$W(n) = 0.54 + 0.46\cos[2\pi n/(N-1)] \tag{C.1}$$

where N is the length of the Hamming window. Generally, the short-time spectral measurements are carried over using a analysis window of size 20-25 ms such that the frames are overlapping with their centers being only 10 ms apart. After windowing, discrete Fourier transform (DFT) is used to convert the windowed speech frame to its frequency domain representation as given in Eqn. C.2 [144].

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{\frac{-j2\pi nk}{N}} \quad \text{for } 0 \leq k \leq N-1 \tag{C.2}$$

where N is the number of points used to compute the DFT. The real and imaginary components of the short-term speech spectrum are squared and added to get the short-term power spectrum as:

$$P(\omega) = Re[S(\omega)]^2 + Im[S(\omega)]^2 \tag{C.3}$$

(ii) **Critical Band Analysis:** Using critical bands for speech analysis is important from the point of hearing mechanism. Critical band for a given center frequency is defined to be the smallest band of frequencies around it which activates the same part of the basilar membrane of the ear.

Consecutive tones lying in the same critical band do not increase the perceived loudness over that of the single tone if they have all the same sound pressure. Therefore critical bandwidth is used to represent the ear's resolving power for simultaneous tones. After finding the power spectrum $P(\omega)$ it is warped along its frequency axis ω into the Bark frequency Ω by [155]:

$$\Omega(\omega) = 6\ln\left[\frac{\omega}{1200\pi} + \left[\left[\frac{\omega}{1200\pi}\right]^2 + 1\right]^{0.5}\right] \tag{C.4}$$

where ω is the angular frequency in rad/s. The resulting warped power spectrum is convolved with the power spectrum of the simulated critical-band [156] masking curve $\Psi(\Omega)$ using Eqn. C.5.

$$\Psi(\Omega) = \begin{cases} 0 & \text{for } \Omega < -1.3, \\ 10^{2.5(\Omega+0.5)} & \text{for } -1.3 \leq \Omega \leq -0.5, \\ 1 & \text{for } -0.5 < \Omega < 0.5, \\ 10^{-1.0(\Omega-0.5)} & \text{for } 0.5 \leq \Omega \leq 2.5, \\ 0 & \text{for } \Omega > 2.5 \end{cases} \tag{C.5}$$

The discrete convolution of $\Psi(\Omega)$ with $P(\omega)$ yields samples of the critical-band power spectrum as:

$$\Theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i)\Psi(\Omega) \tag{C.6}$$

This convolution reduces the spectral resolution of $\Theta(\Omega)$ in comparison with the original $P(\omega)$. The Bark scale spectrum $\Theta(\Omega)$ is then down-sampled by resampling every one Bark. Typically, 18 spectral samples of $\Theta[\Omega(\omega)]$ are used to cover the 0-16.9 Bark (0-5 kHz) analysis bandwidth in 0.994 Bark steps.

(iii) **Equal Loudness Pre-emphasis:** The sampled $\Theta[\Omega(\omega)]$ is pre-emphasized by the simulated equal loudness curve given in Eqn. C.7.

$$E[\Omega(\omega)] = E(\omega)\Theta[\Omega(\omega)] \tag{C.7}$$

The function $E(\omega)$, adopted from [157], simulates the non-equal sensitivity of human hearing at about 40dB level and is given by:

$$E(\omega) = [(\omega^2 + 56.8 \times 10^6)\omega^4]/[(\omega^2 + 6.3 \times 10^6)^2 \times (\omega^2 + 0.38 \times 10^9)] \quad \text{(C.8)}$$

The last equation is the transfer function of a filter with asymptotes of 12dB/oct between 0 and 400 Hz, 0dB/oct between 400 and 1200 Hz, 6dB/oct between 1200 and 3100 Hz, and 0dB/oct between 3100 Hz and the Nyquist frequency. This approximation is well up to 5000 Hz. This equation is used to simulate hearing resolution power.

Finally, the first (0 Bark) and the last (Nyquist Frequency) samples are not well defined and so, are made equal to the values of their nearest neighbors. Thus, $E[\Omega(\omega)]$ begins and ends with two equal-valued samples.

(iv) **Intensity Loudness Power-Law:** To approximate the power law of hearing [158], cubic root amplitude compression is applied after the PLP filterbank.

$$\Phi(\Omega) = E(\Omega)^{0.33} \quad \text{(C.9)}$$

This operation simulates the nonlinear relation between the intensity of sound and its perceived loudness. This also allows low order all-pole modeling because, together with the psychophysical equal-loudness pre-emphasis, it reduces the spectral amplitude variation of the critical band spectrum. Low model order is necessary to reduce the computational cost.

(v) **Autoregressive Modeling:** Here, $\Phi(\Omega)$ is approximated by the spectrum of an all-pole model using the autocorrelation method of an all-pole spectral modeling. Inverse DFT (IDFT) is applied to $\Phi(\Omega)$ to yield the autocorrelation function dual to $\Phi(\Omega)$. The first $M+1$ autocorrelation values are used to solve the Yule-Walker equations for the autoregressive coefficients of the M^{th} order all-pole model.

(vi) **Cepstral Coefficients:** The perceptual linear predictive cepstral coefficient (PLPCC) are computed from the PLP coefficients obtained in the last step using the recursion formula [119] as :

$$c_1 = a_1 \quad \text{(C.10)}$$

$$c_n = \sum_{k=1}^{n-1}(1 - k/n)a_k c_{n-k} + a_n \ , \quad 1 < n < M \tag{C.11}$$

The cepstral coefficients described so far are referred to as the static features which capture only the average frequency distribution for a given frame. The extra information about the temporal dynamics of the signal is obtained by computing the first order derivatives (delta coefficients) and the second order derivatives (delta-delta coefficients) of the static cepstral coefficients [148–150] as :

$$\Delta c_m(n) = \frac{\sum_{i=-T}^{T} k_i c_m(n+i)}{\sum_{i=-T}^{T} |i|} \tag{C.12}$$

where $c_m(n)$ denotes the m^{th} feature for the n^{th} time frame, k_i is the i^{th} weight and T is the number of successive frames used for computation. Generally, T is taken as 2.

D

Perceptual-MVDR Cepstral Coefficients

Contents

D.1 Perceptual Minimum Variance Distortionless Response (PMVDR) Cepstral Coefficient Computation 160

D.1 Perceptual Minimum Variance Distortionless Response (PMVDR) Cepstral Coefficient Computation

The basic idea behind the perceptual minimum variance distortionless response (PMVDR) cepstral coefficient referred to as 'PMVDR' features [127] is to use the minimum variance distortionless response (MVDR) [129] spectral estimator and directly perform the warping of the discrete Fourier transform (DFT) power spectrum rather than using a filterbank based processing.

The step-by-step description of the PMVDR feature extraction process is given below:

(i) The speech signal is first pre-emphasized using a filter having a transfer function as:

$$H(z) = 1 - az^{-1} \tag{D.1}$$

where the value of a controls the slope of the filter and is usually between 0.9 to 1.0. The pre-emphasized speech signal is then divided into frames of length N and weighted by an analysis window, usually a Hamming window of size 20-25 ms. The short-time spectral measurements are carried over such that the frames are overlapping with their centers being only 10 ms apart. After windowing, DFT is used to convert the windowed speech frame to its frequency domain representation as given in Eqn. D.2 [144].

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nk}{N}} \; ; \qquad 0 \leq k \leq N-1 \tag{D.2}$$

where N is the number of points used to compute the DFT. The real and imaginary components of the short-term speech spectrum are squared and added to get the short-term power spectrum as:

$$P(\omega) = Re[S(\omega)]^2 + Im[S(\omega)]^2 \tag{D.3}$$

(ii) The entire 2π warped frequency range, to which the DFT power spectrum is to be scaled, is divided into N equi-spaced points. This gives the N linearly spaced spectral points over the warped frequency space.

$$\hat{\omega}[i] = 2i\pi/N \; , \qquad i = 0, ..., N-1 \tag{D.4}$$

(iii) The linear frequencies and the DFT indices that correspond to these warped frequencies are calculated as:

$$\omega[i] = \tan^{-1} \frac{(1-\alpha^2)\sin(\hat{\omega}[i])}{(1+\alpha^2)\cos(\hat{\omega}[i]) + 2\alpha} , \quad i = 0, ..., N-1 \tag{D.5}$$

$$\hat{k}[i] = \frac{\omega[i]N}{2\pi} , \quad i = 0, ..., N-1 \tag{D.6}$$

(iv) An interpolation of the nearest linear spectral values is performed to obtain the warped spectral value

$$k_l[i] = \min(N-2, \hat{k}[i]) , \quad i = 0, ..., N-1 \tag{D.7}$$

$$k_u[i] = \max(1, k_l[i]+1) , \quad i = 0, ..., N-1 \tag{D.8}$$

$$\hat{\mathrm{S}}[i] = (k_u[i] - \hat{k}[i])\mathrm{S}[k_l[i]] + (\hat{k}[i] - k_l[i])\mathrm{S}[k_u[i]] \tag{D.9}$$

where, $k_l[i]$ is the lower nearest linear DFT bin, $k_u[i]$ is the nearest upper linear DFT bin and $\hat{\mathrm{S}}[i]$ is the value of the warped power spectrum that corresponds to DFT bin i. Thus, the spectral value $\hat{\mathrm{S}}[i]$, at the warped frequency index $\hat{k}[i]$, is computed as the linear interpolation of nearest upper, $\mathrm{S}[k_u[i]]$, and lower, $\mathrm{S}[k_l[i]]$, spectral values in the linear frequency space.

(v) The "perceptual autocorrelations lags" are computed by taking the inverse DFT of the "perceptually warped" power spectrum $\hat{\mathrm{S}}$. A p^{th} order linear prediction (LP) analysis is then performed to obtain p LP coefficients via LevinsonDurbin recursion using the perceptual autocorrelation lags [159, 160].

(vi) The p^{th} order MVDR spectrum for all frequencies is computed in a parametric form from the LP coefficients a_i using Eqn. D.10 [128, 161].

$$\mu(k) = \begin{cases} \frac{1}{p_e}\sum_{i=0}^{p-k}(p+1-k-2i)a_i a_{i+k}^* , & k = 0, ..., p \\ \mu^*(-k) , & k = -p, ..., -1 \end{cases} \tag{D.10}$$

(vii) The final cepstrum coefficients are obtained using the straightforward DFT-based approach [144]. In this approach, after obtaining the MVDR coefficients from the perceptually warped spectrum $\hat{\mathrm{S}}$, we take the DFT of the parametrically expressible MVDR spectrum. After taking log, we ap-

ply inverse DFT (IDFT) to return back to the cepstral domain. The resulting first M coefficients are the M static PMVDR cepstral coefficients.

The dynamic PMVDR cepstral coefficients which capture the extra information about the temporal dynamics of the signal are obtained by computing the first order derivatives (delta coefficients) and the second order derivatives (delta-delta coefficients) of the static PMVDR cepstral coefficients as [148–150]:

$$\Delta c_m(n) = \frac{\sum_{i=-T}^{T} k_i c_m(n+i)}{\sum_{i=-T}^{T} |i|} \tag{D.11}$$

where $c_m(n)$ denotes the m^{th} feature for the n^{th} time frame, k_i is the i^{th} weight and T is the number of successive frames used for computation. Generally, T is taken as 2.

Bibliography

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

[2] E. Zwicker, "Subdivision of the audible frequency range into critical bands," *J. Acoust. Soc. Amer.*, vol. 33, no. 2, p. 248, February 1961.

[3] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, July 1997.

[4] O. Scharenborg, "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," *Speech Communication*, vol. 49, no. 5, pp. 336–347, May 2007.

[5] D. C. Burnett and M. Fanty, "Rapid unsupervised adaptation to children's speech on a connected-digit task," in *Proc. ICSLP*, vol. 2, Philadelphia, PA, October 1996, pp. 1145–1148.

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

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

[8] M. Russell, C. Brown, A. Skilling, R. Series, J. Wallace, B. Bonham, and P. Barker, "Applications of automatic speech recognition to speech and language development in young children," in *Proc. ICSLP*, vol. 1, Sydney, Australia, October 1996, pp. 176–179.

[9] A. Hagen, B. Pellom, and R. Cole, "Children's speech recognition with application to interactive books and tutors," in *Proc. ASRU Workshop*, St. Thomas, VI, December 2003, pp. 186–191.

[10] D. Giuliani, O. Mich, and M. Nardon, "A study on the use of a voice interactive system for teaching English to Italian children," in *Proc. ICALT*, Athens, Greece, July 2003, pp. 376–377.

[11] A. Hagen, B. Pellom, S. V. Vuuren, and R. Cole, "Advances in children's speech recognition within an interactive literacy tutor," in *Proc. HLT/NAACL*, Boston, MA, May 2004, pp. 25–28.

[12] A. Alwan, Y. Bai, M. Black, L. Casey, M. Gerosa, M. Heritage, M. Iseli, B. Jones, A. Kazemzadeh, S. Lee, S. Narayanan, P. Price, J. Tepperman, and S. Wang, "A system for technology based assessment of language and literacy in young children: The role of multiple information sources," in *Proc. International Workshop on Multimedia Signal Processing*, Crete, Greece, October 2007, pp. 26–30.

[13] L. Bell, J. Boye, J. Gustafson, M. Heldner, A. Lindström, and M. Wirén, "The Swedish NICE corpus - spoken dialogues between children and embodied characters in a computer game scenario," in *Proc. Interspeech*, Lisbon, Portugal, September 2005, pp. 2765–2768.

[14] L. Bell and J. Gustafson, "Children's convergence in referring expressions to graphical objects in a speech-enabled computer game," in *Proc. Interspeech*, Antwerp, Belgium, August 2007, pp. 2209–2212.

[15] J. Gustafson, L. Bell, J. Boye, A. Lindström, and M. Wirén, "The NICE fairy-tale game system," in *Proc. 5^{th} SIGdial Workshop on Discourse and Dialogue*, Boston, MA, April-May 2004, pp. 23–26.

[16] 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 and Language*, vol. 14, no. 2, pp. 161–175, April 2000.

[17] J. E. Beck, P. Jia, and J. Mostow, "Automatically assessing oral reading fluency in a computer tutor that listens," *Technology, Instruction, Cognition and Learning*, vol. 2, pp. 61–81, 2004.

[18] J. Mostow, S. F. Roth, A. G. Hauptmann, and M. Kane, "A prototype reading coach that listens," in *Proc. Twelfth National Conference on Artificial Intelligence*, Seattle, WA, August 1994, pp. 785–792.

[19] G. A. Korsah, J. Mostow, M. B. Dias, T. M. Sweet, S. M. Belousov, M. F. Dias, and H. Gong, "Improving child literacy in Africa: Experiments with an automated reading tutor," *Information Technologies and International Development*, vol. 6, no. 2, pp. 1–19, Summer 2010.

[20] P. Cosi, R. Delmonte, S. Biscetti, R. A. Cole, B. Pellom, and S. van Vuren, "Italian literacy tutor: tools and technologies for individuals with cognitive disabilities," in *Proc. InSTIL/ICALL Symposium*, Venice, Italy, June 2004, pp. 207–214.

[21] O. Mich, D. Giuliani, and M. Gerosa, "Parling, a CALL system for children," in *Proc. InSTIL/ICALL Symposium*, Venice, Italy, June 2004, pp. 169–172.

[22] R. Cole, D. Massaro, J. de Villiers, B. Rundle, K. Shobaki, J. Wouters, M. Cohen, J. Beskow, P. Stone, P. Connors, A. Tarachow, and D. Solcher, "New tools for interactive speech and language training: Using animated conversational agents in the classrooms of profoundly deaf children," in *Proc. ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education*, London, UK, April 1999, pp. 45–52.

[23] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, "PEAKS - A system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, no. 5, pp. 425–437, May 2009.

[24] E. L. Higgins and M. H. Raskind, "Speech recognition-based and automaticity programs to help students with severe reading and spelling problems," *Annals of Dyslexia*, vol. 54, no. 2, pp. 365–392, 2004.

[25] M. Wald, "An exploration of the potential of automatic speech recognition to assist and enable receptive communication in higher education," *ALT-J: Research in Learning Technology*, vol. 14, no. 1, pp. 9–20, March 2006.

[26] D. J. Feil-Seifer, M. P. Black, M. J. Matarić, and S. Narayanan, "Toward designing interactive technologies for supporting research in autism spectrum disorders," in *Proc. International Meeting for Autism Research*, vol. 1, Chicago, IL, May 2009.

[27] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *J. Acoust. Soc. Amer.*, vol. 105, no. 3, pp. 1455–1468, March 1999.

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

[29] A. Potamianos, S. Narayanan, and S. Lee, "Analysis of children's speech: duration, pitch and formants," in *Proc. Eurospeech*, Rhodes, Greece, September 1997, pp. 473–476.

[30] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, no. 10-11, pp. 847–860, October-November 2007.

[31] M. Benzeghiba, R. D. Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10-11, pp. 763–786, October-November 2007.

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

[33] M. Iseli, Y.-L. Shue, and A. Alwan, "Age- and gender-dependent analysis of voice source characteristics," in *Proc. ICASSP*, vol. 1, Toulouse, France, May 2006, pp. 389–392.

[34] B. Weinrich, B. Salz, and M. Hughes, "Aerodynamic measurements: Normative data for children ages 6:0 to 10:11 years," *Journal of Voice*, vol. 19, no. 3, pp. 326–339, July 2004.

[35] S. Eguchi and I. J. Hirsh, "Development of speech sounds in children," *Acta Oto-Laryngol. Suppl.*, vol. 257, pp. 1–51, 1969.

[36] R. D. Kent, "Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies," *Journal of Speech and Hearing Research*, vol. 19, pp. 421–447, September 1976.

[37] M. Benzeguiba, R. D. Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and intrinsic speech variation," in *Proc. ICASSP*, vol. 5, Toulouse, France, May 2006, pp. 1021–1024.

[38] L. Qun and M. Russell, "Why is automatic recognition of children's speech difficult?" in *Proc. Eurospeech*, Aalborg, Denmark, September 2001, pp. 2671–2674.

[39] R. D. Kent and L. L. Forner, "Speech segment durations in sentence recitations by children and adults," *Journal of Phonetics*, vol. 8, no. 2, pp. 157–168, April 1980.

[40] M. Gerosa, S. Lee, D. Giuliani, and S. Narayanan, "Analyzing children's speech: An acoustic study of consonants and consonant-vowel transition," in *Proc. ICASSP*, Toulouse, France, May 2006, pp. 393–396.

[41] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Amer.*, vol. 87, no. 2, pp. 820–857, February 1990.

[42] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[43] D. G. Childers, "Glottal source modeling for voice conversion," *Speech Communication*, vol. 16, no. 2, pp. 127–138, February 1995.

[44] C. Gobl, "A preliminary study of acoustic voice quality correlates," *STL-QPSR*, vol. 30, no. 4, pp. 9–22, 1989.

[45] I. Karlsson, "Glottal waveform parameters for different speaker types," *STL-QPSR*, vol. 29, no. 2-3, pp. 61–67, 1988.

[46] A. Potamianos and S. Narayanan, "Spoken dialog systems for children," in *Proc. ICASSP*, Seattle, WA, May 1998, pp. 197–201.

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

[48] J. G. Wilpon and C. N. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. ICASSP*, Atlanta, GA, May 1996, pp. 349–352.

[49] M. Eskenazi, "KIDS: A database of children's speech," *J. Acoust. Soc. Amer.*, vol. 100, no. 4, Part 2, pp. 2759–2759, December 1996.

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

[51] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *Proc. Eurospeech*, Rhodes, Greece, September 1997, pp. 2371–2374.

[52] D. Elenius and M. Blomberg, "Comparing speech recognition for adults and children," in *Proc. FONETIK*, Stockholm, Sweden, May 2004, pp. 156–159.

[53] M. Gerosa and D. Giuliani, "Preliminary investigations in automatic recognition of English sentences uttered by Italian children," in *Proc. ESCA ETRW NLP and Speech Technologies in Advanced Language Learning Systems Symposium*, Venice, Italy, June 2004, pp. 9–12.

[54] S. D'Arcy and M. Russell, "A comparison of human and computer recognition accuracy for children's speech," in *Proc. Interspeech*, Lisbon, Portugal, September 2005, pp. 2197–2200.

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

[56] A. Potamianos and R. Rose, "On combining frequency warping and spectral shaping in HMM based speech recognition," in *Proc. ICASSP*, vol. 2, Munich, Germany, April 1997, pp. 1275–1278.

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

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

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

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

[61] X. Cui and A. Alwan, "Adaptation of children's speech with limited data based on formant-like peak alignment," *Computer Speech and Language*, vol. 20, pp. 400–419, July 2006.

[62] S. Umesh, R. Sinha, and S. V. B. Kumar, "An investigation into front-end signal processing for speaker normalization," in *Proc. ICASSP*, vol. 1, Montreal, Canada, May 2004, pp. 345–348.

[63] F. Hönig, G. Stemmer, C. Hacker, and F. Brugnara, "Revising perceptual linear prediction (PLP)," in *Proc. Interspeech*, Lisbon, Portugal, September 2005, pp. 2997–3000.

[64] J. Gustafson and K. Sjölander, "Voice transformations for improving children's speech recognition in a publicly available dialogue system," in *Proc. ICSLP*, Denver, CO, September 2002, pp. 297–300.

[65] Q. Li and M. Russell, "An analysis of the causes of increased error rates in children's speech recognition," in *Proc. ICSLP*, Denver, CO, September 2002, pp. 2337–2340.

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

[67] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multi-variant Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 291–298, April 1994.

[68] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, April 1995.

[69] D. Giuliani, M. Gerosa, and F. Brugnara, "Improved automatic speech recognition through speaker normalization," *Computer Speech and Language*, vol. 20, no. 1, pp. 107–123, January 2006.

[70] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP*, Philadelphia, PA, October 1996, pp. 1137–1140.

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

[72] O. Siohan, T. A. Myrvoll, and C.-H. Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation," *Computer, Speech and Language*, vol. 16, no. 1, pp. 5–24, January 2002.

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

[74] P. Cosi, "On the development of matched and mismatched Italian children's speech recognition systems," in *Proc. Interspeech*, Brighton, UK, September 2009, pp. 540–543.

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

[76] T. Cincarek, I. Shindo, T. Toda, H. Saruwatari, and K. Shikano, "Development of preschool children subsystem for ASR and Q&A in a real-environment speech-oriented guidance task," in *Proc. Interspeech*, Antwerp, Belgium, August 2007, pp. 1469–1472.

[77] S. D'Arcy and M. Russell, "A comparison of human and computer recognition accuracy for children's speech," in *Proc. Interspeech*, Lisbon, Portugal, September 2005, pp. 2197–2199.

[78] M. Gerosa, D. Giuliani, and F. Brugnara, "Speaker adaptive acoustic modeling with mixture of adult and children's speech," in *Proc. Interspeech*, Lisbon, Portugal, September 2005, pp. 2193–2196.

[79] ——, "Towards age-independent acoustic modeling," *Speech Communication*, vol. 51, no. 6, pp. 499–509, June 2009.

[80] K. Shobaki, J.-P. Hosom, and R. Cole, "The OGI kid's speech corpus and recognizers," in *Proc. ICSLP*, vol. 4, Beijing, China, October 2000, pp. 258–261.

[81] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, and M. Wong, "The PF_STAR children's speech corpus," in *Proc. Interspeech*, Lisbon, Portugal, September 2005, pp. 2761–2764.

[82] R. Nisimura, A. Lee, H. Saruwatari, and K. Shikano, "Public speech-oriented guidance system with adult and child discrimination capability," in *Proc. ICASSP*, vol. 1, Montreal, Canada, May 2004, pp. 433–436.

[83] X. Li, Y.-C. Ju, L. Deng, and A. Acero, "Efficient and robust language modeling in an automatic children's reading tutor system," in *Proc. ICASSP*, vol. 4, Honolulu, Hawaii, April 2007, pp. 193–196.

[84] M. Eskenazi and G. Pelton, "Pinpointing pronunciation errors in children's speech: examining the role of the speech recognizer," in *Proc. PMLA Workshop*, Estes Park, CO, September 2002, pp. 48–52.

[85] A. Hagen, B. Pellom, and R. Cole, "Highly accurate children's speech recognition for interactive reading tutors using subword units," *Speech Communication*, vol. 49, no. 12, pp. 861–873, December 2007.

[86] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. ICASSP*, San Diego, CA, March 1984, pp. 42.11.1–42.11.4.

[87] 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. ICASSP*, Detroit, MI, May 1995, pp. 81–84.

[88] "Open source software from the speech group", Wavesurfer version 1.8.5. Online: http://www.speech.kth.se/software/, accessed on 10 Jan 2008.

[89] S. J. Young, G. Evermann, M. J. F. Gales, D. Kershaw, G. Moore, J. J. Odell, D. G. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book version 3.4*. Cambridge, U.K.: Cambridge University Engineering Department, 2006.

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

[91] M. Slaney, "Auditory Toolbox. Version 2," Interval Research Corporation, Tech. Rep. 1998-010, 1998.

[92] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASRU*, Paris, France, September 2000, pp. 181–188.

[93] "BEEP Dictionary". Online: http://mi.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html, accessed on 15 Oct 2010.

[94] J. P. Cabral and L. C. Oliveira, "Pitch-synchronous time-scaling for high-frequency excitation regeneration," in *Proc. Interspeech*, Lisbon, Portugal, September 2005, pp. 1513–1516.

[95] I. Arroabarren and A. Carlosena, "Glottal source parameterization: a comparative study," in *Proc. ITRW VOQUAL*, Geneva, Switzerland, August 2003, pp. 29–34.

[96] A. Andreou, T. Kamm, and J. Cohen, "Experiments in vocal tract normalization," in *Proc. CAIP Workshop: Frontiers in Speech Recognition II*, Piscataway, NJ, July-August 1994.

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

[98] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proc. DARPA Workshop on Speech Recognition*, February 1986, pp. 93–99.

[99] B. Milner and X. Shao, "Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction," *IEEE Trans. Audio Speech Language Processing*, vol. 15, no. 1, pp. 24–33, January 2007.

[100] H. Singer and S. Sagayama, "Pitch dependent phone modelling for HMM based speech recognition," in *Proc. ICASSP*, vol. 1, San Francisco, CA, March 1992, pp. 273–276.

[101] G. Garau and S. Renals, "Combining spectral representations for large vocabulary continuous speech recognition," *IEEE Trans. Audio Speech Language Processing*, vol. 16, pp. 508–518, March 2008.

[102] P. C. Mahalanobis, "On the generalised distance in statistics," in *Proceedings National Institute of Science, India*, vol. 2, no. 1, April 1936, pp. 49–55.

[103] J. L. Miller, "Effects of speaking rate on segmental distinctions," *Perspectives On The Study Of Speech*, pp. 39–74, 1981.

[104] J. Miller and F. Grosjean, "How the components of speaking rate influence perception of phonetic segments," *Journal of Experimental Psychology: Human Performance and Perception*, vol. 7, no. 1, pp. 208–215, 1981.

[105] Q. Summerfield, "Articulatory rate and perceptual constancy in phonetic perception," *Journal of Experimental Psychology: Human Performance and Perception*, vol. 7, pp. 1074–1095, 1981.

[106] M. A. Siegler and R. M. Stern, "On the effects of speech rate in large vocabulary speech recognition systems," in *Proc. ICASSP*, vol. 1, Detroit, MI, May 1995, pp. 612–615.

[107] N. Mirghafori, E. Fosler, and N. Morgan, "Towards robustness to fast speech in ASR," in *Proc. ICASSP*, vol. 1, Atlanta, GA, May 1996, pp. 335–338.

[108] N. Morgan, E. Fosler, and N. Mirghafori, "Speech recognition using on-line estimation of speaking rate," in *Proc. Eurospeech*, Rhodes, Greece, September 1997, pp. 2079–2082.

[109] F. Martínez, D. Tapias, and J. Álvarez, "Towards speech rate independence in large vocabulary continuous speech recognition," in *Proc. ICASSP*, vol. 2, Seattle, WA, May 1998, pp. 725–728.

[110] G. E. Peterson and I. Lehiste, "Duration of syllable nuclei in English," *J. Acoust. Soc. Amer.*, vol. 32, no. 6, pp. 693–703, June 1960.

[111] H. Kuwabara, "Acoustic properties of phonemes in continuous speech for different speaking rate," in *Proc. ICSLP*, vol. 4, Philadelphia, PA, October 1996, pp. 2435–2438.

[112] H. Wakita, "Normalization of vowels by vocal-tract length and its application to vowel identification," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-25, pp. 183–192, April 1977.

[113] T. Pfau, R. Faltlhauser, and G. Ruske, "A combination of speaker normalization and speech rate normalization for automatic speech recognition," in *Proc. ICSLP*, vol. 4, Beijing, China, October 2000, pp. 362–365.

[114] A. M. Sulter and H. P. Wit, "Glottal volume velocity waveform characteristics in subjects with and without vocal training, related to gender, sound intensity, fundamental frequency, and age," *J. Acoust. Soc. Amer.*, vol. 100, no. 5, pp. 3360–3373, November 1996.

[115] D. G. Childers, K. Wu, D. M. Hicks, and B. Yegnanarayana, "Voice conversion," *Speech Communication*, vol. 8, no. 2, pp. 147–158, June 1989.

[116] M. J. Hunt, "Spectral signal processing for ASR," in *Proc. ASRU Workshop*, vol. 1, Keystone, CO, December 1999, pp. 17–25.

[117] L. Gu and K. Rose, "Split-band perceptual harmonic cepstral coefficients as acoustic features for speech recognition," in *Proc. Interspeech*, Aalborg, Denmark, September 2001, pp. 583–586.

[118] M. Jelinek and J.-P. Adoul, "Frequency-domain spectral envelope estimation for low rate coding of speech," in *Proc. ICASSP*, Phoenix, AZ, March 1999, pp. 253–256.

[119] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, no. 6, pp. 1304–1312, June 1974.

[120] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293–309, February 1967.

[121] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis for voiced speech," *J. Acoust. Soc. Amer.*, vol. 47, pp. 634–648, February 1970.

[122] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, April 1990.

[123] J. Psutka, L. Müller, and J. V. Psutka, "Comparison of MFCC and PLP parameterizations in the speaker independent continuous speech recognition task," in *Proc. Eurospeech*, Aalborg, Denmark, September 2001, pp. 1813–1816.

[124] P. Pujol, S. Pol, C. Nadeu, A. Hagen, and H. Bourlard, "Comparison and combination of features in a hybrid HMM/MLP and a HMM/GMM speech recognition system," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 1, pp. 14–22, January 2005.

[125] D.-S. Kim, S.-Y. Lee, and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 1, pp. 55–69, January 1999.

[126] C.-P. Chen, J. Bilmes, and D. P. W. Ellis, "Speech feature smoothing for robust ASR," in *Proc. ICASSP*, vol. 1, Philadelphia, PA, March 2005, pp. 525–528.

[127] U. H. Yapanel and J. H. L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Communication*, vol. 50, no. 2, pp. 142–152, February 2008.

[128] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 221–239, May 2000.

[129] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, August 1969.

[130] S. Dharanipragada, U. H. Yapanel, and B. D. Rao, "Robust feature extraction for continuous speech recognition using the MVDR spectrum estimation method," *IEEE Trans. Audio Speech Language Processing*, vol. 15, no. 1, pp. 224–234, January 2007.

[131] M. J. F. Gales, D. Y. Kim, P. C. Woodland, H. Y. Chan, D. Mrva, R. Sinha, and S. E. Tranter, "Progress in the CU-HTK broadcast news transcription system," *IEEE Trans. Audio Speech Language Processing*, vol. 14, no. 5, pp. 1513–1525, Spetember 2006.

[132] Tsung-hsueh Hsieh and Jeih-weih Hung, "Speech feature compensation based on pseudo stereo codebooks for robust speech recognition in additive noise environments," in *Proc. Interspeech*, Antwerp, Belgium, August 2007, pp. 242–245.

[133] S. Wang, Y.-H. Lee, and A. Alwan, "Bark-shift based nonlinear speaker normalization using the second subglottal resonance," in *Proc. Interspeech*, Brighton, UK, September 2009, pp. 1619–1622.

[134] B. Pellom, "Sonic: The University of Colorado Continuous Speech Recognizer," University of Colorado, Boulder, Colorado, Tech. Rep. TR-CSLR-2001-01, March 2001.

[135] S. Chakroborty, A. Roy, S. Majumdar, and G. Saha, "Capturing complementary information via reversed filter bank and parallel implementation with MFCC for improved text-independent speaker identification," in *Proc. ICCTA*, Kolkata, India, March 2007, pp. 463–467.

[136] W. Yutai, L. Bo, J. Xiaoqing, L. Feng, and W. Lihao, "Speaker recognition based on dynamic MFCC parameters," in *Proc. Image Analysis and Signal Processing*, Taizhou, April 2009, pp. 406–409.

[137] H. Lei and E. Lopez, "Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition," in *Proc. Interspeech*, Brighton, UK, September 2009, pp. 2323–2326.

[138] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99–109, 1943.

[139] S. Umesh, A. Zolnay, and H. Ney, "Implementing frequency-warping and VTLN through linear transformation of conventional MFCC," in *Proc. Interspeech*, Lisbon, Portugal, September 2005, pp. 269–272.

[140] S. Panchapagesan and A. Alwan, "Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC," *Computer Speech and Language*, vol. 23, no. 1, pp. 42–64, January 2009.

[141] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communication*, vol. 20, no. 1-2, pp. 151–173, November 1996.

[142] S. Matsoukas and R. Schwartz, "Improved speaker adaptation using speaker dependent feature projections," in *Proc. ASRU Workshop*, St. Thomas, VI, December 2003, pp. 273–278.

[143] J. W. Picone, "Signal modeling techniques in speech recognition," *Proc. IEEE*, vol. 81, no. 9, pp. 1215–1247, September 1993.

[144] A. V. Oppenheim and R. W. Schafer, *Discrete-time Signal Processing.* Englewood Cliffs, NJ, USA: Prentice-Hall, 1989.

[145] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Amer.*, vol. 8, no. 3, pp. 185–190, January 1937.

[146] J. R. Deller, J. H. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals.* Upper Saddle River, NJ, USA: Prentice Hall PTR, 1993.

[147] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *J. Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, November 2001.

[148] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[149] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 29, no. 3, pp. 342–350, June 1981.

[150] J. S. Mason and X. Zhang, "Velocity and acceleration features in speaker recognition," in *Proc. ICASSP*, vol. 5, Toronto, Canada, April 1991, pp. 3673–3676.

[151] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, no. 4, pp. 532–556, April 1976.

[152] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[153] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[154] B.-H. Juang and L. R. Rabiner, "The segmental K-means algorithm for estimating the parameters of hidden Markov models," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 38, no. 9, pp. 1639–1641, September 1990.

[155] M. R. Schroeder, *Recognition of Complex Acoustic Signals, Life Sciences Research Report 5.* Abakon Verlag, Berlin: edited by T.H. Bullock, 1977.

[156] H. Fletcher, "Auditory patterns," *Review of Modern Physics*, vol. 12, pp. 47–65, 1940.

[157] J. Makhoul and L. Cosell, "LPCW: An LPC vocoder with linear predictive spectral warping," in *Proc. ICASSP*, vol. 1, Philadelphia, PA, April 1976, pp. 466–469.

[158] S. S. Stevens, "On the psychopysical law," *Psychological Review*, vol. 64, pp. 153–181, 1957.

[159] J. Makhoul, "Linear prediction: a tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[160] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Trans. Signal Processing*, vol. 39, no. 2, pp. 411–423, February 1991.

[161] B. Musicus, "Fast MLM power spectrum estimation from uniformly spaced correlations," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 33, no. 5, pp. 1333–1335, October 1985.

List of Publications

Refereed Journals

Manuscripts Published:

1. Shweta Ghai and Rohit Sinha, "Exploring the Effect of Differences in the Acoustic Correlates between Adult's and Children's Speech in context of Automatic Speech Recognition," *EURASIP Journal on Audio, Speech and Music Processing*, vol. 2010, Article ID 318785, 15 pages, 2010. doi:10.1155/2010/318785.

Manuscripts Under Review:

1. Shweta Ghai and Rohit Sinha, "Pitch Normalization by Mel Filterbank Modification for Children's Automatic Speech Recognition," under review for *Speech Communication*

2. Shweta Ghai and Rohit Sinha, "Adaptive Truncation of MFCC Feature for Improving Mismatched ASR of Children's Speech," under review for *IEEE Transactions on Audio, Speech and Language Processing.*

Refereed International/National Conferences

1. Shweta Ghai and Rohit Sinha, "A Study on the Effect of Pitch on LPCC and PLPC Features for Children's ASR in comparison to MFCC," in *Proc. Interspeech*, Florence, Italy, August 2011, pp. 2589–2592.

2. Shweta Ghai and Rohit Sinha, "Enhancing Children's Speech Recognition under Mismatched Condition by Explicit Acoustic Normalization," in *Proc. Interspeech*, Makuhari, Japan, September 2010, pp. 522–525.

3. Shweta Ghai and Rohit Sinha, "Analyzing Pitch Robustness of PMVDR and MFCC Features for Children's Speech Recognition," in *Proc. IEEE SPCOM*, Bangalore, India, July 2010, pp. 1–5.

4. Rohit Sinha and Shweta Ghai, "On the Use of Pitch Normalization for Improving Children's Speech Recognition," in *Proc. Interspeech*, Brighton, UK, September 2009, pp. 568–571.

5. Shweta Ghai and Rohit Sinha, "Exploring the Role of Spectral Smoothing in context of Children's Speech Recognition," in *Proc. Interspeech*, Brighton, UK, September 2009, pp. 1607–1610.

6. Shweta Ghai and Rohit Sinha, "Maximum Likelihood Pitch Normalization for Improving Children's Speech," in *Proc. Fifteenth National Conference on Communications 2009 (NCC 2009)*, Guwahati, India, January 2009, pp. 311-315.

7. Shweta Ghai and Rohit Sinha, "An investigation into the effect of pitch transformation on children speech," in *Proc. IEEE TENCON*, Hyderabad, India, November 2008, pp. 1–6.