IMPROVING CHILDREN'S MISMATCHED ASR THROUGH ADAPTIVE PITCH COMPENSATION

SYED SHAHNAWAZUDDIN

IMPROVING CHILDREN'S MISMATCHED ASR THROUGH ADAPTIVE PITCH COMPENSATION

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

SYED SHAHNAWAZUDDIN

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781039, ASSAM, INDIA

AUGUST 2016

Certificate

This is to certify that the thesis entitled "**Improving Children's Mismatched ASR Through Adaptive Pitch Compensation**", submitted by **Syed Shahnawazuddin** (Roll No. 10610209), a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under my supervision and guidance. The thesis has fulfilled all the requirements as per the regulations of the institute and in my opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated:
Guwahati.

Prof. Rohit Sinha
Professor
Dept. of Electronics and Electrical Engg.
Indian Institute of Technology Guwahati
Guwahati - 781 039, Assam, India.

To

My father **Mr. S. M. S. Ahmad**

&

My mother **Mrs. K. Shamim**

Acknowledgments

This thesis would not have been possible without the immense help and support of several people in various measures. I take this opportunity to convey my most sincere acknowledgments to all of them. I express my deepest and most sincere gratitude to my thesis supervisor Prof. 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 grateful to Prof. P. K. Bora, the chairman of my doctoral committee and to the other members the committee, Prof. S. R. M. Prasanna, Dr. V. Vijaya Saradhi and Dr. Tony Jacob, for providing their valuable suggestions on my work throughout the research. I would also like to thank other faculty members of the Department of Electronics and Electrical Engineering, IIT Guwahati, for their care and support.

I sincerely thank to Prof. S. Umesh, IIT Madras, and Dr. Samudravijaya K, TIFR Mumbai for their valuable suggestions on my work during the project review meetings and workshops, which helped me a lot in my research. I would like to acknowledge the other project members at IIT Madras, IIIT Hyderabd and TIFR Mumbai as well.

I would like to express my sincere gratitude to Mr. Sanjib Das and Dr. L. N. Sharma, Scientific officers, for their enormous help whenever required. I am also grateful to all other technical and office staff members of the department and the computer center for their help. I am very much thankful to my seniors in the EMST Lab Dr. Shweta Ghai, Dr. Govind D., Dr. Sumitra Shukla, Dr. Debadatta Pati, Dr. S. R. Nirmala, Mr. O. P. Singh, Mr. Ramesh Mishra, Mr. Rajib Jana, Mr. Malaya Kumar Nath and Mr. Sunil Y. for their help and support.

I take this opportunity to extend my gratitude towards Dr. Haris B. C., my senior as well as a friend, for his constant help and support during the five years we shared the space together at IITG. I am quite sure, I wouldn't be writing this thesis had he not been there all along.

I am thankful to my friends Deepak K. T, K. Ramesh, Biswajit Dev Sarma, Sumit Shukla,

Abhinav Mishara, Hemant Kumar Kathania, Kukil Khanikar, Santosh Kumar Yadav, Nagaraj Adiga, Banriskhem K Khonglah, Anurag Singh, Sibasankar Padhy, Jiss J., Rohan Kumar Das, Bhanu Priya and Swati Banerjee and all other members in the EMST Laboratory. I am also grateful to the members of the project group with whom I worked especially Anirudhha Deka and Abhishek Dey. I also thank all my juniors especially Nagendra Kumar, Ganji Sreeram and Patri Sathya Karthik.

Above all, I am deeply grateful to my parents, my siblings and my wife. It would not have been possible for me to complete my PhD without their love, support and sacrifice.

Syed Shahnawazuddin

Abstract

With the progress made in the speech processing over the last few decades, an increasing number of user applications employing automatic speech recognition (ASR) systems are being developed. In such human-machine interaction (HMI) applications, the ASR system is often accessed by both the adults and the children. It is well known that the ASR systems trained on the adults' speech exhibit a severely degraded recognition performance when used for transcribing the speech data from the child speakers and vice-versa. One of the ways to achieve good ASR performance for both the adults and the children is to pool a large amount of data from both the group of speakers in the training of the system. The scarcity of the children's speech corpus makes this approach infeasible. On the other hand, pooling a limited amount of children's data with the adults' training set is not found to be very effective. Consequently, this thesis explores the possibility of achieving improved recognition performance for the children's speech on adults' speech trained ASR systems.

To enhance the performance of the children's mismatched ASR, the thesis begins with an exploration of some of the existing adaptation/normalization techniques. Despite the observed improvements with the application of the existing approaches, a large gap still remains between the adults' matched and children's mismatched testing cases. This gap in the performance is attributed to the severe mismatch in the acoustic and the linguistic correlates for the adults' and the children's speech. Among the various sources of mismatch identified in literature, the differences in the size of the vocal organs and the pitch (fundamental frequency) are known to be the most dominant ones. The frequency-warping-based vocal tract length normalization (VTLN) approach is already noted to be very effective in mitigating the ill effects of the differences in the vocal tract dimensions. Therefore, we have tried to analyze the cause and the extent of the pitch-induced mismatch between adults and children in this work. Based on our analysis, we have devised techniques that target the pitch variation across the speakers.

One of the propositions in this thesis is the reduction of the pitch-induced variations through a structured projection of the front-end features and the parameters of the acoustic model to a lower dimensional subspace. Additionally, a spectral smoothing approach is proposed which address the pitch-induced distortions prior to computation of the acoustic features. Both these approaches are found to be highly effective in the context of the children's mismatched ASR. Furthermore, the proposed techniques are noted to result in additive improvements when combined with some of the existing feature-space normalization as well as the model-space adaptation techniques. In order to reduce the latency in the implementation of the model-space adaptation approaches, we have also developed a few fast adaptation techniques suitable for those ASR tasks involving HMI.

Most of the presented techniques were initially developed for the acoustic modelling employing the Gaussian-mixture-based hidden Markov models (GMM-HMM). But, the observed improvements are also found to hold largely for the recently introduced acoustic modelling techniques based on the subspace GMM (SGMM) and the deep neural network (DNN). In the case of the SGMM- and the DNN-based systems, we have also studied the relative effectiveness of the VTLN and the feature-space maximum likelihood linear regression (fMLLR). The fMLLR-based feature normalization is already reported to be very effective for the DNN-based system. On the other hand, the VTLN is observed to be largely ineffective in those cases where the number hidden layers is very large. On the contrary, our study finds that the VTLN is effective not only for the shallow networks but also for the deeper ones in the context of the children's mismatched ASR.

Keywords: Automatic speech recognition, acoustic mismatch, children's ASR, sparse representation, fast adaptation, pitch-adaptive MFCCs, adaptive-liftering, SGMM, DNN.

Contents

List of Figures

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

List of Tables

List of Acronyms

ANN	Artificial neural network
ASR	Automatic speech recognition
BIBO	Bounded-input bounded-output
BN	Bottleneck
BW	Binary-weighting
CAT	Cluster adaptive training
CD	Context-dependent
CDF	Cumulative distribution function
CMLLR	Constrained maximum likelihood linear regression
CMN	Cepstral mean normalization
CNN	Convolutional neural networks
CVN	Cepstral variance normalization
DBN	Deep belief network
DCT	Discrete cosine transform
DFT	Discrete Fourier transform
DNN	Deep neural network
EM	Expectation-maximization
EV	Eigenvoices
Fbank	Filterbank
fMLLR	Feature-space maximum likelihood linear regression
GMM	Gaussian mixture model
HLDA	Heteroscedastic linear discriminant analysis
HMI	Human-machine interaction
HMM	Hidden Markov model
IDFT	Inverse discrete cosine transform
JFA	Joint factor analysis
KEV	Kernel eigenvoices
K-L	Kullback-Leibler
LARS	Least angle regression
Lasso	Least absolute shrinkage and selection operator
LDA	Linear discriminant analysis
LM	Language model
LPCC	linear prediction cepstral coefficients
MAP	Maximum *a posteriori*
MFCC	Mel-frequency cepstral coefficients
ML	Maximum likelihood

MLED	Maximum likelihood eigen-decomposition
MLLR	Maximum likelihood linear regression
MLLT	Maximum likelihood linear transform
MLP	Multi-layer perceptron
MP	Matching pursuit
NMF	Non-negative matrix factorization
OMP	Orthogonal matching pursuit
OOV	Out of vocabulary
PCA	Principal component analysis
PLDA	Probabilistic linear discriminant analysis
PLP	Perceptual linear prediction
PLPCC	Perceptual linear prediction cepstral coefficients
PMVDR	Perceptual minimum variance distortionless response
PRI	Percentage relative improvement
RBM	Restricted Boltzmann machine
RC	Real cepstrum
RMI	Reference model interpolation
RMP	Regression-based model prediction
RNN	Recurrent neural network
RSW	Reference speaker weighting
SA	Speaker adapted
SAT	Speaker adaptive training
SD	Speaker dependent
SGMM	Subspace Gaussian mixture model
SI	Speaker independent
SMAP	Structural MAP
SR	Sparse representation
SSW	Support speaker weighting
STC	Semi-tied covariance
STFT	Short-time Fourier transform
SVM	Support vector machine
SW	Soft-weighting
UBM	Universal background model
VTLN	Vocal-tract length normalization
WER	Word error rate

1

Introduction

Contents

1. Introduction

Automatic speech recognition (ASR) refers to the task of generating the word level transcriptions for the spoken inputs with the use of a computer. Speech forms the most preferred means of communication among humans. One of the prime objectives of the research and development in ASR domain is to achieve effective human-machine interaction (HMI). ASR is considered as the most challenging task among various speech processing tasks, but it finds a growing number of practical of applications. Initially, in the early 20th century, the pace of research in ASR was slow. With the introduction of the hidden Markov model (HMM) [1, 2] into speech recognition (in the 1970s), the research in ASR has seen a tremendous growth. The HMMs have been investigated extensively and emerged as the most dominant technique for acoustic modelling in ASR systems. After successfully addressing simple connected and isolated word recognition tasks, the current focus is on the continuous speech recognition. Further, the size of the recognition vocabulary has also increased from few tens of words to several thousands of words these days.

A simplified block diagram representing the structure of an ASR system is shown in Figure 1.1. During the data acquisition module, the input sound pressure wave is converted into an electrical signal through a microphone and then recorded in digit format. The input raw speech waveforms are first chopped into short-duration frames which are then converted into a suitable parametric representation in the feature extraction module. The chosen parametric representation intends to capture the relevant information in speech signal while removing the redundancies to achieve a compact representation. These short-time parametric representations are also referred to as the acoustic feature vectors. The Mel-frequency cepstral coefficients (MFCC) [3] and the perceptual linear prediction cepstral coefficients (PLPCC) [4] are the two dominant examples of the commonly used ones. A brief summary of those acoustic feature extraction techniques is given in Appendix A.

For acoustic modelling, ASR systems typically employ the HMM-based generative models. In the HMM-based ASR systems, the observations for any particular acoustic unit are assumed to be generated by a finite state machine with state-specific probabilistic distributions. Depending on the complexity of the recognition task, the acoustic units may be chosen as words or sub-words (phone, syllable). At each time unit, change in the state of the system occurs with a certain probability. An observation is generated with some probability distribution whenever a state is entered. One of the ways to model the state observation probabilities is to use a multivariate Gaussian mixture model (GMM). Over the years, efficient algorithms for the training of the parameters of the GMM-HMM system as well as for performing search and decoding

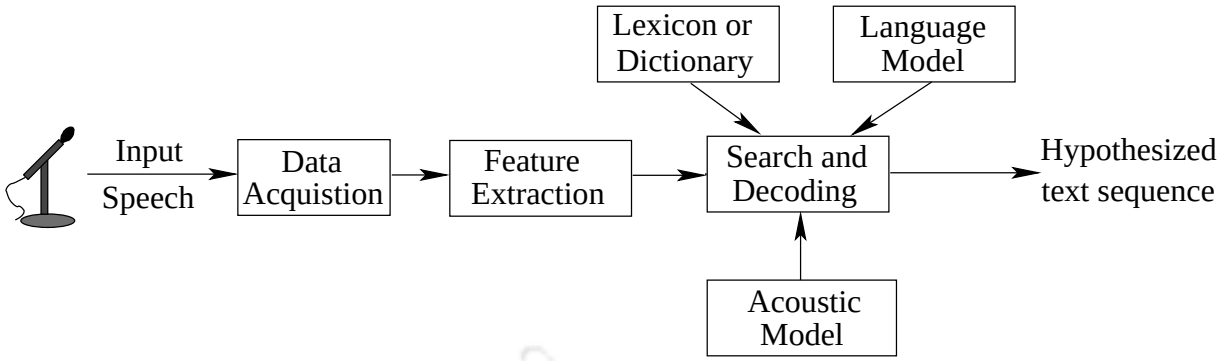

Figure 1.1: A simplified block diagram representing the structure of a typical automatic speech recognition system.

over the network have been developed. The subspace GMM (SGMM) [5] is another acoustic modelling technique reported in literature. In the recent past, there has been a paradigm shift whereby the state-specific posterior probabilities are generated through the neural networks. Deep neural network (DNN) containing many non-linear hidden layers and a very large output layer are being used for this purpose [6]. In Appendix B, a discussion on the three acoustic modelling techniques explored in this thesis are discussed.

The language model (LM) gives the prior probability of a hypothesized sequence of N words, $\widetilde{W} = W_1, W_2, \ldots, W_N$. The LM can be factorised into a product of conditional probabilities as follows:

$$P(\widetilde{W}) = \Pi_{n=1}^{N} P(W_k \,|\, W_{n-1}, \ldots, W_1) \tag{1.1}$$

where the n^{th} word in the sequence is denoted by W_n. The language model is a discrete probability learned on a large text corpus. The n-gram statistical language models are widely used in ASR with bigram and trigram being the most common denominations. The lexicon or the dictionary is used to map sub-word units to actual words present in the vocabulary and the language model.

In the practical ASR systems, the speech data from a large number of speakers is pooled to train the statistical acoustic models. This is done in order to robustly model the speaker, the session and the linguistic variability. Such ASR systems are referred to as the *speaker independent* (SI) systems as the trained acoustic models represent all the speakers in an average sense. For any speaker in particular, the recognition performance can be significantly improved by training an ASR system using the speech data specific to that speaker. The so developed ASR system is called the *speaker dependent* (SD) system. Developing an SD system is often impractical as a huge amount of training data is required for a robust estimation of the acoustic

model parameters. In addition to that, the performance of the developed SD system may not be optimal for other speakers. Several techniques have been devised to map the SI system parameters to the corresponding SD space of a particular speaker using a small amount of data from the target speaker. This approach is termed as *speaker adaptation*, and the developed ASR systems are referred to as the *speaker adapted* (SA) systems. Adapting an SI system to better represent a particular speaker using a small amount of adaptation data is reported to outperform the SI system. A large number of adaptation techniques have been proposed over the last two decades. An excellent review of these techniques is available in [7].

1.1 Motivation for the Study

Over the past few decades, the scope of ASR has spread across a wide range of applications. Several toolkits [8–10] have also been developed to aide the research in ASR and to enable the development of user-specific applications involving speech recognition. Many of those applications involve human-machine interactions. Some examples of such applications are speech-based information retrieval, speech-based web search, reading tutors, language learning tools and entertainment [11–16]. In such tasks, the employed ASR system is required to deal with speakers of both the genders and belonging to various age groups. In general, the SI systems developed for the adult speakers show a degraded recognition performance when used for transcribing the children's speech data. Similar degradation is observed when children's speech trained SI systems are employed for transcribing the adults' speech data. The in-vocabulary word error rate in speech recognition for children's speech is double of that for the adults' speech as reported in [17]. Further, a large degradation in ASR performance is observed when children's speech is tested against the acoustic models derived from adults' speech [18–20]. The salient factors causing the differences in the adults' and children's speech are summarized in [21] as follows *"...differences are attributed mainly to the anatomical and the morphological differences in the vocal-tract geometry, less precise control of the articulators and a less refined ability to control suprasegmental aspects such as prosody."* In addition to that, children's speech has higher fundamental and formant frequencies and greater spectral variability as reported in [22, 23].

Separate acoustic models, one trained using adults' speech and the other using children's speech, can be employed to address this issue. Another way to overcome the aforementioned problem is to pool a large amount of data from both the adult and the child speakers for learning the model parameters. Unfortunately, publicly available children's speech data is very scarce which limits the development of ASR system employing state-of-the-art techniques

unlike the adults' case. Alternatively, one can explore ways of improving the recognition of children's speech on adults' speech trained acoustic models. In this thesis, the latter approach is resorted to and is referred to as the *children's mismatched ASR*. On the other hand, the task of recognizing adults' and children's speech on the ASR systems trained using their respective domain data is referred to as the *matched ASR*.

The work presented in this thesis explores the ways and means to improve the recognition of children's speech on acoustic models trained using adults' speech. One way to achieve this is to improve the trained acoustic models by employing model-space-based adaptation approaches. As mentioned earlier, a number of such techniques have already been reported in literature [7]. In general, the effectiveness of any adaptation technique depends on the number of SI parameters optimally modified with respect to the test speaker/utterance. This, in turn, depends on the amount of available adaptation data. The conventional techniques such as the maximum *a posteriori* (MAP) adaptation [24] and the maximum likelihood linear regression (MLLR) adaptation [25] work very well when the available adaptation data is large enough. At the same time, these approaches become largely ineffective in those cases where the adaptation data is small (5-10 seconds only). The fast adaptation techniques, on the other hand, try to modify the system parameters with such small amount of adaptation data. In such approaches, the basic premise is that the adapted model parameters lie in a low-dimensional subspace spanned by a set bases (known beforehand) [7].

Another way to improve the recognition performance of the mismatched ASR system is through the use of feature-space normalization techniques like the vocal tract length normalization (VTLN) [26] and the constrained MLLR (CMLLR) [27]. Due to a large mismatch in the acoustic characteristic of the children's and adults' speech, the use of normalization techniques become quite effective [28]. Moreover, the feature-space normalization techniques can be employed in combination with the model-space adaptation approaches. The combination of the two is observed to result in significant improvements in the recognition performance of the mismatched ASR system [28–30].

Despite the use of existing adaptation/normalization approaches, a large gap still exists in the recognition performances for the acoustically matched and mismatched cases. In the following, we discuss some of the major factors responsible for the observed degradations. Once the major mismatch factors are figured out, adaptation/normalization techniques specific to children's mismatched ASR can be developed.

1.2 Challenges in Children's ASR

Over the last few decades, recognition of adults' speech has witnessed significant improvement. On the other hand, limited efforts have been made towards improving children's ASR. Large differences in both the acoustic and the linguistic correlates between the speech from the adult and the child speakers make the automatic recognition of children's speech much tougher [21, 31–34]. During the growing phase, the ability of a child to produce varying speech sounds properly and accurately improves with the age [35]. Furthermore, as stated earlier, children's have smaller vocal organs compared to the adults. Consequently, speech from child speakers has higher fundamental and formant frequencies and greater spectral variability [22, 23]. In addition to that, the overall speaking rate is slower in the case of children and they have more variability in the speaking rate as well [21, 28]. The children's speech is reported to have greater values for the mean and the variance for the acoustic correlates of speech than those for the adults' [28]. For example, as observed in [22], for most of the vowel phonemes the area of the $F1$-$F2$ formant ellipses is larger for the children than for the adults. Consequently, children's speech suffers from a higher degree of inter- and intra-speaker acoustic variability than the adults' speech [21, 36]. From the linguistic perspective, children are more likely to use *imaginative words*, *ungrammatical phrases* and *incorrect pronunciations* as stated in [16].

To compensate for these sources of mismatch, a number of techniques have been suggested in literature for children's speech recognition under the mismatched conditions. Burnett and Fanty proposed a fast approach for compensating the formant scaling using a speaker-dependent warping of the frequency scale through offsets in the *Bark*-domain [19]. Gustafson and Sjölander reported an improved recognition of children's speech on publicly available fixed adults' speech trained ASR with an explicit reduction of the pitch of the signals [37]. In addition, the improved phone classification with pitch-dependent normalization [38] and speech reconstruction by pitch prediction using the MFCC features have also been reported [39]. Further, the VTLN has been noted to be very effective for the children's mismatched ASR [28]. Lately, it has been shown that the standard Mel-filterbank involved in the MFCC feature extraction is not able to provide sufficient smoothing especially for the high-pitched child speakers [40]. As a result, there can be a significant mismatch in the variances of the higher-order MFCCs for adults' and children's speech [41]. To reduce the resulting mismatch, a simple binary weighting (BW) of the features that essentially truncates some higher-order coefficients in the base MFCC features is explored in [40].

1.3 Objectives and Contributions

The work presented in this thesis mainly deals with the task of recognizing children's speech on acoustic models trained using adults' speech. Gross mismatch in the acoustic attributes between the adult and the child speakers can lead to a severe degradation in the recognition performance as noted in the earlier works [17–21, 28]. On account of the differences in the dimensions of the vocal organs, children's speech exhibit not only much higher formant frequencies but also much higher fundamental frequency values than those observed in the case of adults' speech. Thus, apart from the vocal tract length differences, large differences in the pitch values for the above mentioned group of speakers also contribute to acoustic mismatch.

The fact that the Mel-filterbank involved in the standard MFCC feature extraction fails to smooth out the pitch harmonics effectively for the high-pitched child speakers is already highlighted in [40]. As a result of the pitch-induced distortions in the spectral envelope, a significant mismatch is noted in the variances of the higher-order MFCCs for the adults' and children's speech [41]. In the GMM-HMM-based acoustic models trained on adults' speech, the variances corresponding to the higher-order coefficients are usually much smaller relative to the lower-order ones in all three streams of the MFCC feature vector. On computing the likelihood of children's test data with respect to models trained on adults' speech, the Mahalanobis distance metric happens to enhance the distance score for the higher-order feature coefficients due to a higher precision in the acoustic models [41]. This leads to a degradation in the likelihood which, in turn, degrades the recognition performance.

The thesis aims at reducing the pitch-induced mismatch through the use of model-space-based adaptation approaches as well as feature-space-based normalization techniques. In this regard, we have analyzed the cause and the extent of pitch-induced distortions in the acoustic feature extraction process from the signal processing perspective. This has enabled us to develop an acoustic feature extraction process that is robust to the gross variation of the pitch across the speakers. Moreover, the techniques proposed in this work are intended towards those applications that involve human-machine interactions. Consequently, the latency (or the computational cost) involved in implementing the adaptation/normalization approaches is also given due focus in this thesis. In the presented work, the ASR systems employed in experimental evaluation are trained and optimized for the adult speakers. The developed mismatch reduction approaches are then implemented to improve the recognition of children's speech keeping the overall computational cost as low as possible. Though most of the adaptation/normalization techniques were initially developed keeping the GMM-HMM-based ASR systems into consider-

ation, the developed approaches are found to be effective even in the case of acoustic modelling based on SGMM and DNN.

The salient contributions of the thesis are summarized as follows:

- Exploring the effectiveness of existing model-interpolation-based fast adaptation techniques to improve the recognition performance of the children's mismatched ASR system.
- Developing a sparse representation-based fast adaptation technique having lower computational cost compared to the existing approaches.
- A structured low-rank feature-projection-based fast adaptation approach is developed to address the pitch-dependent mismatch in the variances of the acoustic features.
- Exploring the role of pitch-adaptive signal processing in the extraction of the MFCC features.
- A novel adaptive-liftering-based approach is proposed for deriving pitch-robust MFCC features.
- Further, the effectiveness of the pitch-adaptive MFCCs for children's mismatched ASR is also demonstrated in the context of the SGMM- and the DNN-based acoustic modelings.

1.4 Thesis Organization

The remaining of the thesis is organized as follows:

It begins with a brief survey of the salient model- and feature-domain adaptation/normalization techniques in Chapter 2. The thesis aims at devising low latency adaptation approaches. Keeping the latency low makes the adaptation/normalization techniques amenable for HMI tasks. Consequently, the fast adaptation techniques such as those based on acoustic model interpolation are given greater focus. The description provided in this chapter serves as a ready reference to all those techniques that are referred later in the thesis.

In Chapter 3, the model-interpolation-based fast adaptation techniques using sparse representation are discussed. First, a technique that makes use of sparse coding for the dynamic selection of acoustically close basis models is presented in Section 3.2. This is followed by a discussion on a novel approach that uses sparse coding for adapting the Gaussian mean parameters of a GMM-HMM-based ASR system in Section 3.4. The presented approaches are evaluated on the matched and the mismatched test sets and are found to be quite effective. In addition

to that, the use of sparse coding is observed to reduce the computational cost significantly. Furthermore, the experimental evaluations presented in this chapter also highlight the degradation in the recognition performance that is observed when children's speech is recognized on acoustic models trained using adults' speech. This serves as the motivation for developing fast adaptation approaches specific to the children's mismatched ASR task.

An analysis of the observed degradations in the case of children's mismatched ASR is presented in Chapter 4. The study presented in this chapter highlights the cause of pitch-induced distortions noted when high-pitched signals are analyzed using static front-end signal processing techniques. We make use of the conclusions drawn in this chapter as the theoretical basis for the rest of the work reported in the thesis.

In the case of children's speech, the variances of the higher-order cepstral coefficients is observed to be much greater than that for the adults' speech as mentioned earlier. A fast adaptation technique based on structured low-rank feature projections for normalizing the mismatch in the variances is presented in Chapter 5. Further, a study on effectively combining the low-rank feature projections with different model-space-based adaptation techniques as well as feature-space normalization approaches is also presented in this chapter.

In Chapter 6, we explore the effectiveness of pitch-adaptive signal processing in the front-end speech parameterization process. In this regard, we propose a simple technique based on pitch-adaptive liftering. The proposed as well as the existing pitch-adaptive techniques are found to be quite effective in the case of the children's mismatched ASR task.

The role of some of the adaptation/normalization techniques is explored in the context of the mismatched ASR system employing the DNN-HMM-based acoustic modelling in Chapter 7. In addition to that, the acoustic modelling based on the SGMM is also explored. The use of the low-rank feature projection and the pitch-adaptive acoustic features are shown to be quite effective even in the SGMM- and the DNN-based acoustic modelling for the mismatched ASR task.

Finally, the concluding remarks are given in Chapter 8 while summarizing the salient results and discussing the scope of future work.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

2

Survey of Speaker Adaptation and Normalization

Contents

The adaptation techniques for the GMM-HMM system are generally classified into three families of approaches viz. the maximum *a posteriori* (MAP), the linear transform and the speaker cluster family as given in [7]. The MAP and the linear transform families belong to the class of conventional adaptation techniques. On the other hand, the acoustic model-interpolation-based fast adaptation techniques belong to the speaker-cluster/speaker-space family. In this work, we have further classified the fast adaptation techniques as shown in Figure 2.1. In model-interpolation-based approaches a set of predefined model parameters are linearly combined to derive the adapted model parameters. These techniques have been explored for the adaptation of the Gaussian mean as well as the mixture-weight parameters. In the case of Gaussian mean parameter adaptation, the acoustic models to be interpolated are either fixed *a priori* or determined dynamically for each test speaker/utterance. The performances for the discussed approaches are also subject to the availability of the true transcription of the adaptation data and the information about the speaker.

In ASR systems, the adaptation is usually performed in three ways: *batch*, *incremental* and *instantaneous* modes [42,43]. In the batch mode, full adaptation data is available *a priori* to the system while in the incremental mode, on the other hand, it is made available *dynamically*. In the case of instantaneous adaptation, the given test utterance is used to estimate the adaptation transformation. In the remaining of this thesis, this case will be referred to as the *utterance-specific* mode. Incremental adaptation applies to the case where the system adapts in an unsupervised manner every utterance the user says. In other words, when the n^{th} test utterance is made available, the statistics of all previous $(n-1)$ utterances are accumulated with the current statistics. Depending on the availability of the true transcription of the data, the adaptation is referred to as *supervised* or *unsupervised*.

In this chapter, a brief review of the conventional adaptation approaches is presented in Section 2.1. This is followed by the formal definition of the model-interpolation-based fast adaptation in Section 2.2. A detailed survey of the acoustic model-interpolation-based fast adaptation techniques developed over the years is presented next. We begin with a discussion on the Gaussian mean adaptation techniques in Section 2.3 which is followed by a discussion on approaches based on mixture-weight interpolation in Section 2.4. Those techniques that happen to be an extension of some other earlier reported work are all discussed under one broad heading. This is followed by a discussion on some of the speaker normalization approaches reported in literature in Section 2.5.

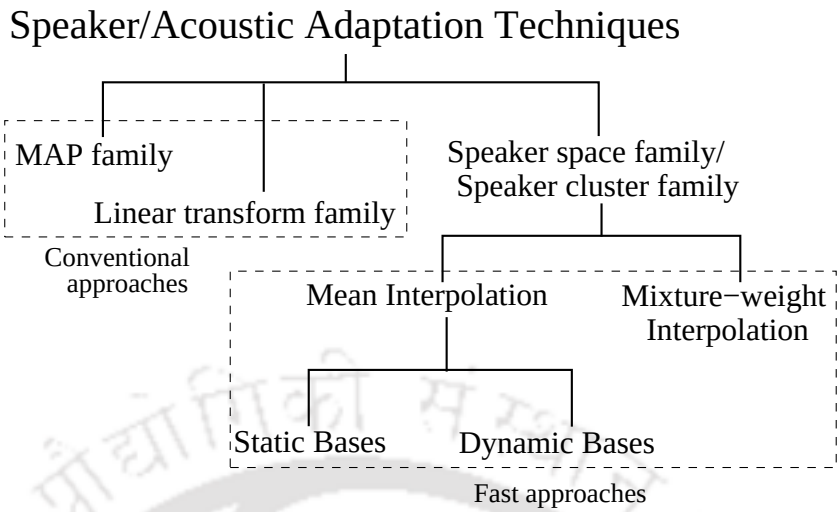

Figure 2.1: Broad classification of different techniques developed for speaker/acoustic adaptation in the context of ASR systems.

2.1 Conventional Adaptation Approaches

2.1.1 MAP family

The maximum *a posteriori* (MAP) technique is the oldest and the most powerful adaptation approach. In this technique, the parameters of a well trained SI model are used as the prior information to estimate the new speaker-specific model parameters. Given the set of GMM-HMM-based acoustic models parameters $\mathbf{\Lambda}$ for the SI system and adaptation data vector $\mathbf{O} = (\mathbf{o}_1, \ldots, \mathbf{o}_L)$ which is a series of L observation sequences, the adapted model parameters are estimated using the following criterion:

$$\hat{\mathbf{\Lambda}} = \arg\max_{\mathbf{\Lambda}} p(\mathbf{\Lambda}|\mathbf{O}) = \arg\max_{\mathbf{\Lambda}} p(\mathbf{O}|\mathbf{\Lambda})\, p(\mathbf{\Lambda}). \tag{2.1}$$

Given the prior mean vector $\bar{\boldsymbol{\phi}}^r$ for the r^{th} Gaussian in the SI system, the estimate of the mean for the adapted system is given by

$$\hat{\boldsymbol{\phi}}^r = \frac{\tau\bar{\boldsymbol{\phi}}^r + \sum_{l=1}^{L}\gamma^r(l)\mathbf{o}_l}{\tau + \sum_{l=1}^{L}\gamma^r(l)} \tag{2.2}$$

where τ is a meta-parameter which controls the bias between the maximum likelihood (ML) and the prior estimate of the mean while $\gamma^r(l)$ is the posterior probability of the data with respect to the r^{th} Gaussian in the SI system for the l^{th} frame. It is clear from the above equation that if the adaptation data increases, the MAP estimate converges to the ML estimate. Similar update rules for the covariance matrices and the mixture-weights can also be derived [24].

The main drawback of the MAP adaptation is that it is a slow a process and only those Gaussians that are observed in the adaptation data get adapted. Several techniques addressing the shortcomings of the MAP adaptation have been proposed in literature [44–46]. Regression-based model prediction (RMP) [45], first updates the means using the standard MAP adaptation. Those parameters that have received a good amount of data for estimation (source parameters) are then used to estimate the values of the poorly adapted parameters whose final value is a linear combination of the initial MAP estimate and the weighted predicted values (weighted in terms of the inverse variance estimates). The RMP is reported to result in better performance than the MAP when the available adaptation data is less. Towards addressing the shortcomings of conventional MAP, another algorithm employing tree-structured constraints in the estimation of parameters and is referred to as the structural MAP (SMAP) [47].

2.1.2 Linear transform family

The transformation of the SI parameters based on the maximum likelihood linear regression (MLLR) belongs to this class of acoustic adaptation. Like the MAP, the MLLR also requires an initial continuous density HMM system. In the MLLR, it is assumed that the differences in speaker characteristics are mainly revealed by the mean vectors of the GMM-HMM -based ASR system. Hence, given the adaptation data, generally the Gaussian means parameters are re-estimated in this technique [25]. Extension of this technique can be used to estimate the covariance matrices as well [48].

In the case of MLLR, for any particular Gaussian r in the unadapted SI systems, the adapted mean vector is modeled as

$$\boldsymbol{\phi}^r = \mathbf{P}\bar{\boldsymbol{\phi}}^r + \mathbf{b} \tag{2.3}$$

where $\mathbf{P}$ is the MLLR transformation matrix and $\mathbf{b}$ is an optional bias vector. Alternatively, we can rewrite (2.3) as

$$\boldsymbol{\phi}^r = \mathbf{W}\boldsymbol{\xi}^r \tag{2.4}$$

where $\mathbf{W} = [\mathbf{b}\ \mathbf{P}]$ is the $D \times (D+1)$ transformation matrix (D being the dimension of the observation vector) and $\boldsymbol{\xi}^r = [1\ \bar{\boldsymbol{\phi}}^{r^T}]^T$ is the extended mean vector. In order to estimate the the transformation matrix $\mathbf{W}$, the likelihood of the adaptation data is maximized using the expectation-maximization (EM) algorithm. In the case of MLLR, the auxiliary function to be

maximized is given by the following equation:

$$Q(\mathbf{\Lambda}, \bar{\mathbf{\Lambda}}) = \sum_{q} \mathcal{F}(\mathbf{O}, \mathbf{q}|\mathbf{\Lambda}) \ \log\left(\mathcal{F}(\mathbf{O}, \mathbf{q}|\bar{\mathbf{\Lambda}})\right) \tag{2.5}$$

where $\mathbf{\Lambda}$ and $\bar{\mathbf{\Lambda}}$ are the current and the updated GMM-HMM parameters, respectively, $\mathbf{O}$ is the adaptation data and $\mathbf{q}$ is the set of all possible state sequences. Usually the transformation matrix is shared by a number of Gaussians if a sufficient amount of adaptation data is not available for the estimation of the transformation matrices for each and every Gaussian [7]. The tying of transformation matrix over a number of Gaussians is performed using regression class tress based clustering. In this case, the Gaussians that are close in the acoustic space are clustered together and are transformed using the same matrix $\mathbf{W}$. The clustered components are then arranged in a tree structure. As the amount of data increases, the tree may be descended to an appropriate depth and the transformation matrix can be estimated precisely for all other Gaussians [49,50]. Due to the imposed constraints, all the Gaussians mean vectors can be updated using the MLLR transformation with a lesser amount of adaptation data than the MAP. This subsequently makes the MLLR approach comparatively faster than the MAP technique.

Even though it had been assumed earlier that the important speaker-specific characteristics are reflected by the Gaussian means, the MLLR can be used to update the covariance matrices as well. Once the mean transformation matrix has been evaluated, the covariance can be estimated in a similar fashion using the following relation as given in [48, 51]

$$\mathbf{C} = \mathbf{H}\bar{\mathbf{C}}\mathbf{H}^{T} \tag{2.6}$$

where $\bar{\mathbf{C}}$ is the covariance matrix for any Gaussian in the SI GMM-HMM system and $\mathbf{H}$ is the variance transform. A closed-form solution for estimating $\mathbf{H}$ using EM algorithm has also been derived in [51]. A variation of MLLR is the constrained-MLLR (CMLLR) [27]. In this technique, unlike the MLLR, same transformation matrices are applied to the Gaussian mean vectors as well as the covariance matrices. The following relations give the r^{th} mean vector and the covariance matrix transformed using the CMLLR based transformation.

$$\boldsymbol{\phi}^{r} = \mathbf{P}_{c}\,\bar{\boldsymbol{\phi}}^{r} + \mathbf{b}_{c} \tag{2.7}$$

$$\mathbf{C}^{r} = \mathbf{P}_{c}\,\bar{\mathbf{C}}^{r}\,\mathbf{P}_{c}^{T}. \tag{2.8}$$

In a practice, CMLLR is implemented as a feature-space normalization technique rather than

model parameter transformation since it can be used to transform the l^{th} observation sequence $\mathbf{o}_l$ as follows

$$\tilde{\mathbf{o}}_l = \mathbf{P}_c^{-1}\mathbf{o}_l + \mathbf{P}_c^{-1}\mathbf{b}_c \tag{2.9}$$

where $\tilde{\mathbf{o}}_l$ is the estimate of the transformed observation vector.

2.2 Model-Interpolation-based Fast Adaptation

One of the simplest fast adaptation technique is the 1-best model-search-based approach. In this case, instead of decoding the test data $\mathbf{O}$ using the SI system $\mathbf{\Lambda}$, a search is performed over a set of predefined acoustic models $\{\mathbf{\Lambda}_i\}$. Generally, $\mathbf{\Lambda}_i$ is derived by modifying the SI parameters using some developmental data. The search for the optimal acoustic model is done by computing the likelihood of the given data with respect each of the models $p(\mathbf{O}|\mathbf{\Lambda}_i)$ and selecting the one which results in the highest likelihood. Consequently, no parameter re-estimation is performed during the adaptation phase. This approach is quite effective when the test data happens to be from a speaker whose data has been used for the training of the acoustic models. In addition to that, the scope of this approach becomes limited if each of the acoustic models does not robustly represent all the phonetic contexts of the target language. The acoustic model-interpolation-based adaptation approaches are an extension of the 1-best search technique.

The model interpolation techniques assume that the adapted model parameters lie in a low (K) dimensional acoustic space. This low dimensional subspace is, in turn, synthesized by a linear interpolation of the set of candidate acoustic models $\{\mathbf{\Lambda}_i\}_{i=1}^{K}$ (bases). If $\boldsymbol{\zeta}_i$ denotes the acoustic model parameter in $\mathbf{\Lambda}$ modified using the developmental data to derive $\mathbf{\Lambda}_i$, the adapted model parameter for the r^{th} Gaussian is given by

$$\boldsymbol{\zeta}^r = \boldsymbol{\zeta}_0^r + \eta_1\boldsymbol{\zeta}_1^r + \ldots + \eta_i\boldsymbol{\zeta}_i^r + \ldots + \eta_K\boldsymbol{\zeta}_K^r \tag{2.10}$$

where $\boldsymbol{\zeta}_0^r$ is an optional bias and $\{\eta_i\}$ is the set of interpolation weights. These interpolation weights are generally global in nature, i.e., the same weights are used for estimating the adapted parameters for all the Gaussians. The remaining parameters in the adapted model are borrowed from the SI system. Due to the imposed constrains, only the interpolation weight parameters need to be estimated for each test speaker/utterance leading to a huge reduction in the computational cost. Furthermore, the reduction in the number of parameters to be estimated implies that those can be robustly estimated using even a small amount of adaptation

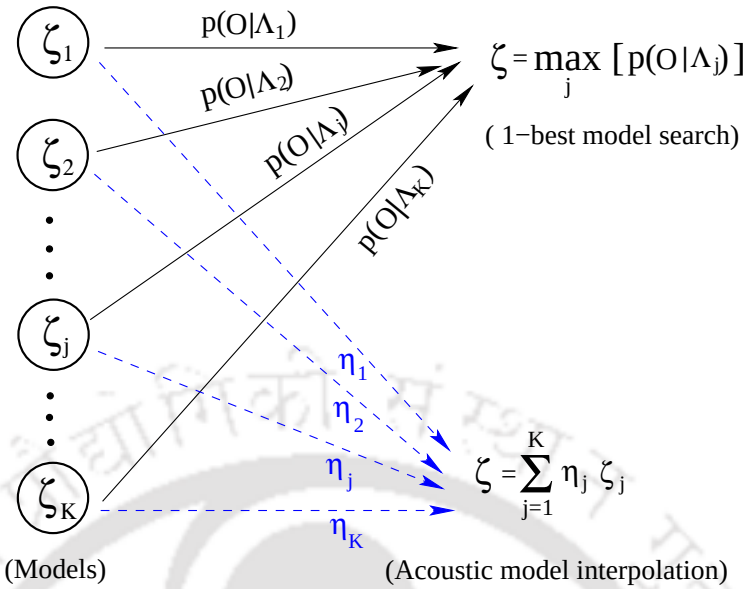

Figure 2.2: Acoustic model-interpolation-based adaptation approach. In this case $\boldsymbol{\zeta}_j$ denotes the model parameters in $\boldsymbol{\Lambda}_j$ to be interpolated to derive the adapted model parameter $\boldsymbol{\zeta}$ and η_j denotes the interpolation weights.

data. The diagrammatic representation of model-interpolation-based fast adaptation approach is outlined in Figure 2.2. It is evident from the figure that the 1-best search is an special case where a unity weight is assigned to the most likely model and zero to the rest.

In the case of model-interpolation-based approaches, there are two main problems to be addressed.

(i) How are the bases (candidate acoustic models) derived ?

(ii) How are the interpolation weights estimated ?

Usually, the SA models corresponding to each of the speakers in the training dataset serve as the candidate models. Clustering of the training speakers to obtain cluster-specific acoustic models is another way to derive the bases. These SA/cluster models are, in turn, obtained through the MAP or the MLLR transformation of the SI model parameters. In the works reported in literature, the SA models are derived by adapting only the Gaussian means or the mixture-weight parameters of the SI system. The model-interpolation-based approaches have been explored for the adaptation of the Gaussian means and the mixture-weights only. In other words, $\boldsymbol{\zeta}_i^r$ and $\boldsymbol{\zeta}^r$ in (2.10) denote either the Gaussian mean vector or the mixture-weight. Recently, a number of techniques have been reported that employ a dynamic selection of bases prior to model interpolation. This approach is found to outperform the earlier techniques that

employed fixed bases. In the case of dynamic selection, a set of acoustically close models are selected using the given adaptation data. The selected model parameters are then interpolated to synthesize the adapted model parameters. The dynamic selection further reduces the number of parameters to be estimated since a small subset is selected from all the SA models. In all cases of model interpolation, the weights are estimated using an iterative ML approach. Some techniques, especially in the case of mixture-weight adaptation, impose the constraints of *non-negativity* and *sum-to-one* on the interpolation weights during estimation.

2.3 Gaussian Mean Interpolation Schemes

As already discussed, the model-interpolation-based approaches have been employed for the Gaussian mean as well as the mixture-weight adaptation. In the following we first discuss the procedure of interpolation weight estimation using the maximum likelihood criterion which is at the core of the majority of the techniques. This is followed by a discussion on some the techniques that resort to model interpolation for the adaptation of Gaussian mean parameters of the HMM.

2.3.1 ML estimation of interpolation weights

The model mean-interpolation-based adaptation techniques use an iterative ML estimation process to derive the interpolation weights. Given the adaptation data observation sequence, the weight parameters are so estimated that the likelihood of the adaptation data with respect to current estimate of the adapted model increases in comparison to that obtained with respect to the previous estimate. This iterative weight estimation process is similar to the generation of the MLLR transformation [25]. Let $\mathbf{O} = (\mathbf{o}_1, \ldots, \mathbf{o}_L)$ be the adaptation data which is a series of L observation sequences, $\gamma^r(l)$ be the posterior probability of occupying the r^{th} Gaussian of a state given that the observation sequence $\mathbf{o}_l$ is generated, $\bar{\mathbf{C}}^r$ be the covariance matrix for the r^{th} Gaussian component in the SI model and D be the dimension of feature vector. Let $\mathbf{\Phi} = [\boldsymbol{\phi}_1 \ldots \boldsymbol{\phi}_i \; ldots \boldsymbol{\phi}_K]$ be the matrix of K bases to be interpolated. The *mean-supervector* $\boldsymbol{\phi}_i$ is derived from the i^{th} SA/cluster model by stacking all the Gaussian mean vectors. The order in which the Gaussian means are concatenated should remain the same in all the supervectors. The adapted Gaussian mean supervector is then modeled as

$$\boldsymbol{\phi} = \mathbf{\Phi}\boldsymbol{\eta} \tag{2.11}$$

where $\boldsymbol{\eta} = [\eta_1 \dots \eta_K]^T$ is the global interpolation weight vector. Note that the bias in (2.10) has been ignored in this case. Let the current and the updated HMM parameters be denoted by $\boldsymbol{\Lambda}$ and $\bar{\boldsymbol{\Lambda}}$, respectively. The *auxiliary function* $Q(\boldsymbol{\Lambda}, \bar{\boldsymbol{\Lambda}})$ to be maximized (ignoring the terms independent of $\boldsymbol{\phi}$) is expressed as per the following:

$$Q(\boldsymbol{\Lambda}, \bar{\boldsymbol{\Lambda}}) = -\frac{1}{2} P(\mathbf{O}|\boldsymbol{\Lambda}) \sum_{r=1}^{R} \sum_{l=1}^{L} \gamma^r(l) \Big(\mathbf{o}_l - \boldsymbol{\Phi}^r \boldsymbol{\eta} \Big)^T (\bar{\mathbf{C}}^r)^{-1} \Big(\mathbf{o}_l - \boldsymbol{\Phi}^r \boldsymbol{\eta} \Big) \tag{2.12}$$

where $\boldsymbol{\Phi}^r$ corresponds to the r^{th} component in the matrix $\boldsymbol{\Phi}$ and R is total number of Gaussians in the HMM. Maximizing (2.12) with respect to $\boldsymbol{\eta}$ gives the updated estimate of the interpolation weights.

$$\begin{aligned} & \frac{d}{d\boldsymbol{\eta}} Q(\boldsymbol{\Lambda}, \bar{\boldsymbol{\Lambda}}) = 0 \\ \text{or} \quad & -\frac{1}{2} P(\mathbf{O}|\boldsymbol{\Lambda}) \frac{d}{d\boldsymbol{\eta}} \left[\sum_{r=1}^{R} \sum_{l=1}^{L} \gamma^r(l) \Big(\mathbf{o}_l - \boldsymbol{\Phi}^r \boldsymbol{\eta} \Big)^T (\bar{\mathbf{C}}^r)^{-1} \Big(\mathbf{o}_l - \boldsymbol{\Phi}^r \boldsymbol{\eta} \Big) \right] = 0 \\ \Longrightarrow \quad & \boldsymbol{\eta} = \left[\sum_{r=1}^{R} \left(\sum_{l=1}^{L} \gamma^r(l) \right) \boldsymbol{\Phi}^{r^T} (\bar{\mathbf{C}}^r)^{-1} \boldsymbol{\Phi}^r \right]^{-1} \quad \left[\sum_{r=}^{R} \boldsymbol{\Phi}^{r^T} (\bar{\mathbf{C}}^r)^{-1} \left(\sum_{l=1}^{L} \gamma^r(l) \mathbf{o}_l \right) \right] \end{aligned} \tag{2.13}$$

Using the updated value of $\boldsymbol{\eta}$, an estimate of the adapted model mean supervector is then given as follows:

$$\hat{\boldsymbol{\phi}} = \sum_{i=1}^{K} \eta_i \, \boldsymbol{\phi}_i \tag{2.14}$$

This procedure may be iterated a few times to ensure convergence in ML sense. The detailed steps involved in the ML weight estimation process are given in Algorithm 1. Except the Gaussian means, all the remaining parameters of the adapted model (covariance, mixture-weight and transition matrices) are borrowed from the SI model.

2.3.2 Reference speaker weighting

In the reference speaker weighting (RSW) technique [52], each speaker in the training data for which a reasonably accurate estimate of the centroid (mean) for each of the phonetic classes is available is represented as a reference speaker. To deal with the issues of small amount of adaptation data while creating the reference speakers, adaptation of the monophone models is performed. The Gaussian mean parameters of each phone class are concatenated to represent the reference speaker model. For example, for a speaker s ($s = 1, \dots, N$) and a phonetic class r ($r = 1, \dots, R$), the centroid is represented as a vector $\mathbf{c}_{r,s}$ and each of the R centroids are

Algorithm 1 Estimation of the interpolation weights using the ML criterion and the derivation of the adapted model mean supervector $\boldsymbol{\phi}$

Given: The SI GMM-HMM system $\boldsymbol{\Lambda}$, the adaptation data observation sequence $\{\mathbf{o}_l\}_{l=1}^{L}$ and the matrix of K bases to be interpolated, $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1 \dots \boldsymbol{\phi}_K]$, where $\{\boldsymbol{\phi}_i\}_{i=1}^{K}$ is the set of mean adapted basis supervectors

Step 1: Initialize the weight vector $\boldsymbol{\eta}^{(1)} = [\eta_1^{(1)} \dots \eta_K^{(1)}]^T$

for $j = 1, j \leq \mathrm{J}$, **do**

Step 2: Derive the adapted model mean supervector

$$\boldsymbol{\phi}^{(j)} = \sum_{i=1}^{K} \eta_i^{(j)} \boldsymbol{\phi}_i$$

Step 3: Replace the Gaussian mean vectors in $\boldsymbol{\Lambda}$ with the corresponding values in the adapted mean supervector $\boldsymbol{\phi}^{(j)}$ to derive the adapted model $\boldsymbol{\Lambda}^{(j)}$

if $j = J$ **then**

Break

else

Step 4: Compute the posterior $\gamma^r(l)$ of the adaptation data with respect to the r^{th} Gaussian using adapted model

Step 5: Accumulate the statistics for the weight estimation

$$\mathbf{W} = \sum_{r=1}^{R} \left(\sum_{l=1}^{L} \gamma^r(l) \right) \boldsymbol{\Phi}^{r^T} (\bar{\mathbf{C}}^r)^{-1} \boldsymbol{\Phi}^r$$

$$\mathbf{Z} = \sum_{r=1}^{R} \boldsymbol{\Phi}^{r^T} (\bar{\mathbf{C}}^r)^{-1} \left(\sum_{l=1}^{L} \gamma^r(l)\, \mathbf{o}_l \right)$$

where $\bar{\mathbf{C}}^r$ denotes the covariance matrix corresponding to the r^{th} Gaussian in $\boldsymbol{\Lambda}^{(j)}$

Step 6: Re-estimate the interpolation weight vector

$$\boldsymbol{\eta}^{(j+1)} = \mathbf{W}^{-1}\mathbf{Z}$$

end if

end for

Step 7: Derive the final adapted model mean supervector as

$$\boldsymbol{\phi} = \sum_{i=1}^{K} \eta_i^{(\mathrm{J})} \boldsymbol{\phi}_i$$

then concatenated to provide the supervector-based representation for the speaker **s** as

$$\boldsymbol{\phi}_s = [\mathbf{c}_{1,s}^T \, \mathbf{c}_{2,s}^T \, \ldots \, \mathbf{c}_{R,s}^T]^T. \tag{2.15}$$

The entire set of N reference speakers is represented in form of a matrix as

$$\boldsymbol{\Phi} = [\boldsymbol{\phi}_1 \; \boldsymbol{\phi}_2 \; \ldots \; \boldsymbol{\phi}_N]. \tag{2.16}$$

In testing, the mean supervector of the test speaker/utterance is derived using 2.16 as

$$\boldsymbol{\phi} = \boldsymbol{\Phi}\boldsymbol{\eta}. \tag{2.17}$$

The interpolation weight vector $\boldsymbol{\eta} = [\eta_1 \; \ldots \; \eta_S]^T$ is estimated following a hill climbing procedure subject to the constraints that $\eta_i \geq 0$ and $\sum_{i=1}^{S} \eta_i = 1$. The centroid for the test speaker/utterance corresponding to a particular phonetic class, r, can be individually obtained from the portion of $\boldsymbol{\Phi}$ and $\boldsymbol{\phi}$ representing that phonetic class as given by

$$\boldsymbol{\phi}^r = \boldsymbol{\Phi}^r\boldsymbol{\eta}. \tag{2.18}$$

2.3.3 Cluster adaptive training

In the speaker clustering technique, the training data is used define a set of speaker cluster models where each model represents a cluster of speakers being similar in some sense [7]. The adaptation data is then used to select the best cluster model. The clustered adaptive training (CAT) [53] is an extension of this technique where instead of choosing only the 1-best cluster model, a linear combination of all the cluster models is used. Hence, the mean vector of the r^{th} Gaussian for the test speaker/utterance is a weighted sum of the cluster mean vectors for the same Gaussian as given by (2.14). These interpolation weights are estimated in ML sense following Algorithm 1. Furthermore, in the case of CAT, instead of estimating global weights, a number of Gaussians can be tied into a regression class and one weight parameter is then estimated for each of the V regression classes. In that case, the adapted model mean parameter is given as

$$\boldsymbol{\phi}^r = \boldsymbol{\Phi}^r\boldsymbol{\eta}^v \tag{2.19}$$

where $\boldsymbol{\eta}^v = [\eta_1^v \ldots \eta_K^v]^T$ is the weight vector for the v^{th} regression class ($v = 1, \ldots, V$) and $\boldsymbol{\Phi}^r = [\boldsymbol{\phi}_1^r \ldots \boldsymbol{\phi}_K^r]$ is the matrix of K cluster mean vectors for the r^{th} Gaussian. The bases

interpolation weight vector $\boldsymbol{\eta}^v$ is now computed as

$$\boldsymbol{\eta}^v = (\mathbf{W}^v)^{-1}\mathbf{Z}^v \tag{2.20}$$

where the required statistics now given by

$$\mathbf{W}^v = \sum_{r \in R^v} \left(\sum_{l=1}^{L} \gamma^r(l) \right) \boldsymbol{\Phi}^{r^T} \bar{\mathbf{C}}_r^{-1} \boldsymbol{\Phi}^r \tag{2.21}$$

$$\mathbf{Z}^v = \sum_{r \in R^v} \boldsymbol{\Phi}^{r^T} \bar{\mathbf{C}}_r^{-1} \left(\sum_{l=1}^{L} \gamma^r(l)\, \mathbf{o}_l \right) \tag{2.22}$$

where R^v is the set of Gaussians belonging to the v^{th} regression class. The CAT also provides means to re-estimate the cluster model parameters using the adaptation data. Consequently, the weight estimation and the cluster model parameter re-estimation can be interleaved to perform an adaptive training.

2.3.4 Eigenvoice

In the Eigenvoice (EV) technique [54], the adapted mean parameters are represented as the weighted sum of canonical speaker model parameters. The difference of this approach from RSW and CAT is that these canonical models are obtained by the principal component analysis (PCA) [55] performed on the covariance/correlation matrix constructed from the entire set of N speaker-specific mean supervectors. The mean supervector $\boldsymbol{\phi}_s$ is obtained by concatenating the R Gaussian mean vectors of the SA model corresponding to the s^{th} speaker. Unlike RSW, the SA models are derived from the state-clustered triphone HMMs. After PCA, the K eigenvectors having the largest eigenvalues are retained and these are referred to as the *eigenvoices*. The EVs capture the broad acoustic and linguistic variations present among the speakers in the training set. Consequently, their linear interpolation spans a low dimensional space which is supposed to capture all the speaker variability including those which have not been observed before. If $\mathbf{E}$ denotes the matrix of the extracted K-eigenvoices $\{\mathbf{e}_i\}_{i=1}^K$, the Gaussian mean for a test speaker/utterance is then modeled as a linear interpolation of the K-eigenvoices

$$\boldsymbol{\phi} = \mathbf{e}_0 + \sum_{i=1}^{K} \eta_i\, \mathbf{e}_i \tag{2.23}$$

where $\mathbf{e}_0$ represents the origin of the subspace spanned by the eigenvoices and is generally chosen the SI mean supervector ($\bar{\boldsymbol{\phi}}$) while $\{\eta_i\}_{i=1}^K$ is the set of basis interpolation weights. Given $\mathbf{E}$

and the adaptation data, the interpolation weight vector $\boldsymbol{\eta} = [1\ \eta_1 \dots \eta_K]^T$ is estimated using maximum likelihood eigen-decomposition (MLED) [54] which is a slightly varied version of Algorithm 1 as discussed later in Section 2.3.6.

2.3.5 Variants of Eigenvoice

There are two major drawbacks in the EV adaptation approach, viz., the eigenvoices capture only the inter-speaker variabilities and the performance saturates even when a large amount of adaptation data is available. A number of techniques addressing the saturation in recognition performance have been reported in literature. In all of those approaches, instead of estimating global interpolation weights as done in the case of EV, one interpolation weight parameter is estimated either for each segmental eigenvoice [56] or for each of the feature dimensions [57,58]. Consequently, the number of parameters (interpolation weights) to be estimated for adaptation is large. This, in turn, requires a greater amount of adaptation data for their robust estimation. All the variants of the EV approach are reported to outperform the conventional EV when the adaptation data is significantly large. It is worth noting that for very small amount of adaptation data (adaptation data duration being $\leq$ 6 secs), the performances for all these techniques fall even below that for the unadapted SI system. In the following we provide a brief overview of each of the techniques.

2.3.5.1 Segmental EV

In this approach [56], each of the mean supervector derived from the SA models is first segmented into M sub-supervector before PCA is performed. If, for any speaker s, $\boldsymbol{\phi}_s$ represents the mean supervector of dimension D_s, then the segmented supervector corresponding to that speaker is represented as $\boldsymbol{\phi}_{s,m}$ such that

$$D_s = \sum_{m=1}^{M} D_{s,m} \tag{2.24}$$

where $D_{s,m}$ is the dimension for the m^{th} segment. This segmentation may be done at the Gaussian mixture level or at the model level or even at the speech feature level. PCA is then performed on each of the segmented supervectors to derive M sub-eigenspaces and their corresponding eigenvoices. Let $\boldsymbol{\rho}_{k,m}$ represent the m^{th} segmental eigenvoice where $k = 1, \dots, K$ (total number of eigenvoice retained after PCA). Any new speaker is then represented as a weighted

linear combination of the segmental eigenvoices corresponding to the sub-eigenspaces given by

$$\boldsymbol{\rho}_m = \sum_{i=1}^{K} \eta_{m_i} \boldsymbol{\rho}_{m_i}, \qquad m = 1, \ldots, M \tag{2.25}$$

where η_{m_i} is the i^{th} interpolation weight for the m^{th} segmental eigenvoice. The MLED is employed to estimate the corresponding weights in this approach as well. As as result of the segmentation, the number of weights to be estimated increases by a factor of M, i.e., $K \times M$ weights are estimated instead of K as done in the EV approach.

2.3.5.2 Kernel EV

The Kernel EV (KEV) [59] technique extends the EV approach by introducing the use of non-linear kernel PCA instead of the linear PCA. The speaker mean supervector $\{\boldsymbol{\phi}_s\}_{s=1}^{N}$ corresponding to each of the N speakers is first mapped to high dimensional space using a kernel function $\vartheta(\cdot\,,\cdot)$. Unlike the EV, the kernel induced mapping to a high dimensional space results in the loss of state information in the obtained supervector $\vartheta(\boldsymbol{\phi}_s)$. Consequently, a maximum likelihood estimation of interpolation weights cannot be performed. In order to overcome this, composite kernels are employed. Each of the R Gaussian mean vectors in the speaker-specific supervector is separately mapped using separate kernel functions as

$$\vartheta(\boldsymbol{\phi}_s) = [\vartheta_1(\boldsymbol{\phi}_s^1), \ldots, \vartheta_r(\boldsymbol{\phi}_s^r), \ldots, \vartheta_R(\boldsymbol{\phi}_s^R)]^T \tag{2.26}$$

where $\vartheta^r(\boldsymbol{\phi}_s^r)$ is the kernel function mapping for the r^{th} Gaussian component in the s^{th} speaker-specific mean supervector. With these modifications, the maximum likelihood estimation of interpolation weights becomes possible. Two different kind of composite kernels namely, the *direct sum* and the *tensor product* kernel were explored in that work. The mean supervector for a test speaker/utterance is derived in a manner analogous to that done in the EV approach. The performance evaluation performed on the TIDIGITS speech corpus [60] shows that the KEV approach outperforms the conventional EV technique. The use of kernel mapping increases the computational complexity and therefore this approach is slower than the standard EV.

2.3.5.3 2-D PCA-based EV

In a manner very similar to the segmental EV approach, 2-dimensional (2-D) PCA has been implemented for creating the canonical models in the work reported in [57]. In the conventional PCA-based approach, each speaker is represented as a supervector and covariance/correlation matrix is constructed out these vectors. In 2-D PCA, on the other hand, each speakers is

represented as matrix. If there are R Gaussians in the SA model, then the Gaussian mean of a particular speaker s is represented as a matrix given as follows:

$$\boldsymbol{\phi}_s = [\boldsymbol{\phi}_s^1 \ \ldots \ \boldsymbol{\phi}_s^R] \tag{2.27}$$

where $\boldsymbol{\phi}_s^r \in \mathbb{R}^{D\times 1}$ is the D-dimensional mean vector corresponding to the r^{th} Gaussian. Consequently, like the segmental EV approach, this technique results in D sub-eigenvoices when PCA is performed on the sample covariance matrix created from the rows of the speaker matrices. If K-eigenvoices are extracted for d^{th} dimension in the feature vector $\{\mathbf{e}_{d_k}\}_{k=1}^K$, then $K \times D$ interpolation weights are required to be estimated.

2.3.5.4 PLDA-based EV

In a recent approach for deriving the basis models, the probabilistic linear discriminant analysis (PLDA) has been used in [58]. PLDA-based EV is very similar to the 2-D PCA-based approach. Like the latter, each speaker s is represented as a $R \times D$ matrix given by (2.28).

$$\boldsymbol{\phi}_s = [\boldsymbol{\phi}_s^1 \ \ldots \ \boldsymbol{\phi}_s^R]^T \tag{2.28}$$

Using PLDA, the d^{th} column of $\boldsymbol{\phi}_s$ is modeled as

$$\boldsymbol{\phi}_{d,s} = \boldsymbol{\phi}_{mean} + \mathbf{F}\mathbf{h}_{d,s} + \mathbf{G}\mathbf{w}_{d,s} + \epsilon_{d,s} \tag{2.29}$$

where $\boldsymbol{\phi}_{mean}$ is the mean of the training samples $\mathbf{O}$, $\mathbf{F}$ and $\mathbf{G}$ are the bases of the *between-speaker* and *within-speaker* subspaces, respectively. The co-ordinates of the two spaces are given by $\mathbf{h}_{d,s}$ and $\mathbf{w}_{d,s}$, respectively, while $\epsilon_{d,s}$ denotes the residual noise. Considering only the within-speaker variations for modeling an unknown speaker, the bases for interpolation are the columns of $\mathbf{G} \in \mathbb{R}^{K\times D}$, selecting the number of factors to be K. The estimated adapted model mean parameter is then given as

$$\hat{\boldsymbol{\phi}} = \bar{\boldsymbol{\phi}} + \mathbf{G}\boldsymbol{\eta} \tag{2.30}$$

where $\boldsymbol{\eta} \in \mathbb{R}^{K\times D}$ is the interpolation weight matrix. Therefore, like the 2-D PCA-based approach, $K \times D$ interpolation weights are required to be estimated.

2.3.6 Comparison of CAT and EV

Though the EV and the CAT approaches are very similar, there are the following differences:

(i) Unlike the EV, the bias vector $\boldsymbol{\zeta}_0$ is optional in the case of the CAT.

(ii) In the case of the EV, normalized observation sequence is used in the computation of posterior probabilities during weight estimation. Consequently, the first order statics in the case of the EV $\left(\sum_l \gamma(l)\,[\mathbf{o}_l - \boldsymbol{\zeta}_0]\right.$, where $\boldsymbol{\zeta}_0$ is generally chosen as the corresponding SI mean parameter$\left.\right)$ is different from that for the CAT. When a bias vector is used in the case of the CAT as well, the weight estimation process given in Algorithm 1 becomes exactly same as that of MLED.

(iii) Depending on the amount of adaptation data available, the number of weight classes V can be dynamically determined in the case of the CAT as done in the determination of regression classes in MLLR [49]. Hence, the CAT has provisions for the estimation of weights for each regression class instead of global weights unlike the EV approach. In essence, this is similar to the approaches reported in [56–58]. Consequently, employing regression-class-specific interpolation weights can overcome the saturation issues but at the cost of increased number of parameters.

(iv) The CAT approach also provides means for the re-estimation of cluster model parameters. This helps in dealing with the unseen speakers scenario since the original cluster model parameters are biased towards the seen speakers.

2.3.7 Techniques employing dynamic selection of bases

During the past decade a number of techniques have been reported in literature that intend to modify the RSW approach in order to obtain improved recognition performances. These techniques differ from the RSW approach in the following three ways:

i.) The reference speaker models are derived by the adaptation of SI model developed using state-clustered triphones.

ii.) The bases to be interpolated to derive the adapted model parameters are dynamically selected for each test speaker/utterance.

iii.) The constraints of non-negativity and summing-to-one imposed on the interpolation weight estimation are relaxed.

In the following we discuss some of those techniques that are build over the RSW framework.

2.3.7.1 Support speaker weighting

The support speaker weighting (SSW) [61] technique is an extension of the RSW, in which the dynamic selection of reference speakers/bases is proposed based on support vector machines (SVM). A set reference models is first created by deriving a SA model for each of the N training speakers using MLLR transformation of the SI model. The Gaussian mean parameters from each of the SA models are concatenated to create N supervectors. The test speaker/utterance is also represented by a Gaussian mean supervector extracted from the MAP adapted SI model. A set of dynamic basis is then determined by training a SVM with the supervectors of the reference speakers and that of the test speaker/utterance constituting the two classes. The Gaussian mean parameters of the K numbers of SA models ($K < N$), corresponding to the K supports for the SVM training, are linearly interpolated to obtain the adapted model parameters as given by (2.14). The interpolation weights for the selected bases are estimated in ML sense as per Algorithm 1.

2.3.7.2 Improved reference speaker weighting

Like the SSW, for each of the N speakers in the training data, a SA model is created using the MLLR transformation of mean parameters of the SI model in the improved reference speaker weighting (Im-RSW) [62] approach as well. For each test speaker/utterance, a set of K models are chosen that have the highest likelihood among all the SA models for the given adaptation data. The interpolation weight vector estimation for the dynamically selected reference speaker models is done using the iterative approach given in Algorithm 1. The constraints non-negativity and summing to one are relaxed in this technique as well. The final adapted mean parameters are given by (2.14).

2.3.7.3 Reference model interpolation

In reference model interpolation (RMI) technique [63], first, the interpolation weights are estimated for all the SA models using a single iteration of the ML estimation process. The K number of SA models having the largest magnitude for the interpolation weights are selected for deriving the adapted model. The interpolation weights for these K models are then re-estimated using the ML criterion. In RMI technique, it is also proposed that when a large amount of test data is available, the model parameters for the selected SA models should be adapted with respect to the test data using MAP/MLLR. This, in turn, leads to improvements in the recognition performance when the available adaptation data is significantly large. Unlike

Im-RSW and SSW, a bias vector $\mathbf{b}$ is also used while deriving the final adapted Gaussian mean parameters as given by

$$\boldsymbol{\phi} = \sum_{i=1}^{K} \eta_i \, \boldsymbol{\phi}_i + \mathbf{b} \tag{2.31}$$

This bias vector $\mathbf{b}$ is jointly estimated with the interpolation weights. Like the EV approach, the bias vector is assigned unity weight and hence acts as the origin of the acoustic space spanned by the selected SA models.

2.3.7.4 Sparse representation-based basis selection

Adaptation techniques employing sparse representation (SR) [64] have also been reported recently in literature. A smoothing of ML weights using SR is proposed recently in [65]. In that work, the training data is split into 34-homogeneous sets and a Gaussian mean supervector is derived for each set. The interpolation weights are first derived for all the homogeneous sets using an ML-based estimation procedure. The derived weights are then smoothed by imposing an ℓ_1-regularization criterion using Lasso-based optimization [66]. Consequently, some of the ML weight parameters shrink to zero resulting in a better modeling of the current test environment. A drawback of this technique is that the initialization using the ML weights and then smoothing using the ℓ_1-regularization significantly increases the complexity.

In a similar work reported in [67], SR techniques are used for model-interpolation-based adaptation. Algorithms for SR based on matching pursuit (MP) [68] and ℓ_1-regularized SR [66] are used in that work. Though, employing MP for bases selection is reported to be very fast, comparatively better performances are obtained by using ℓ_1-regularized SR. The projected gradient algorithm [69] is employed for optimization in the case of ℓ_1-regularized SR. The complexity of gradient descent is much more than that of the MP algorithm. In addition to that, the projected gradient algorithm requires to be initialized for optimal convergence. The bases selected using the MP-based approach and the corresponding values of their interpolation weights are used as the initialization for the optimization algorithm. The weights for those atoms that are not selected using the MP are initialized to *zero*. A sequential ML criterion is used during sparse coding for optimization.

2.4 Mixture-Weight Interpolation Scheme

An adaptation technique based on the Gaussian mixture-weight interpolation for deriving the adapted model parameters for the test speaker/utterance is proposed in [70]. This technique

is very similar to the EV approach. First, a small number of latent speaker models (bases) are obtained from the training set using non-negative matrix factorization (NMF). To do the same, a non-negative matrix $\mathbf{A}_w$ is defined wherein each column is the Gaussian mixture-weight supervector corresponding to the training speakers. Just like the PCA, this supervector matrix is then decomposed into two non-negative matrices, $\mathbf{X}_w$ and $\mathbf{Y}_w$, using NMF. Each column of $\mathbf{X}_w$ is a supervector corresponding to one of the latent/reference speaker. While each column of $\mathbf{Y}_w$ gives the linear combination weights that had been used to approximate the corresponding latent speaker supervector using the columns of $\mathbf{A}_w$. The matrix decomposition is optimized by minimizing the distance measure between $\mathbf{A}_w$ and its approximation $\tilde{\mathbf{A}}_w$ (where $\tilde{\mathbf{A}}_w = \mathbf{X}_w\mathbf{Y}_w$). Different distance measures namely mean square error (MSE) [71, 72] and Kullback-Leibler (K-L) divergence criterion [73] are employed for error minimization.

Once the latent models are derived, the adapted Gaussian mixture-weights for the test speaker/utterance are expressed as a linear combination of the supervector of the latent speakers. The interpolation weights for the latent speakers are estimated in the ML sense subject to the condition that the weights are *non-negative* and *sum-to-one*. Given the latent speaker matrix $\mathbf{X}_w$ and the adaptation data, the latent coefficients (interpolation weights) $\mathbf{y}_s$ for a test speaker/utterance are estimated iteratively using the E-M algorithm (similar to that given in Algorithm 1). If there are K latent speakers, then the i^{th} iteration for the k^{th} weight estimation, $\{y_s^{(i+1)}(k)\}_{k=1}^{K}$, is

$$y_s^{(i+1)}(k) = \sum_{l=1}^{L}\sum_{r=1}^{R} \frac{\gamma^r(l)\mathbf{X}_w^r(k)}{\sum_{k=1}^{K}\mathbf{X}_w^r(k)y_s^i(k)} y_s^i(k) \tag{2.32}$$

where $\gamma^r(l)$ is the posterior probability of the adaptation data with respect to r^{th} Gaussian in the current estimate of the adapted model. The Gaussian mixture-weight vector for the adapted model $\boldsymbol{\lambda}$ is then given by

$$\boldsymbol{\lambda} = \mathbf{X}_w\mathbf{y}_s \tag{2.33}$$

Like other model-interpolation-based approaches, since only a small number of parameters are estimated, even a small amount of adaptation data is sufficient. A very similar approach for mixture-weight adaptation is described in [74]. The NMF-based Gaussian mixture-weight interpolation is extended in [75] to support adaptive training using the feature-space MLLR (fMLLR) and the vocal tract length normalization (VTLN). Another extension of the NMF-based Gaussian mixture-weight interpolation is proposed in [76] where the EV based mean adaptation is combined with the mixture-weight adaptation. Additional improvements in recognition performance are reported when the two techniques combined together.

2.5 Acoustic Feature Normalization Techniques

Despite the presence of a number of intra- and inter-speaker differences in the speech units, human beings are able to recognize speech from a large set of speakers having a wide variety of accents and physiological characteristics [45]. Consequently, on can assume that the human brain is capable of performing some kind of normalization that filters out the individual differences. An effort to replicate this attribute of the human brain is the basic premise on which the speaker normalization techniques are based. The main objective of these techniques is to define a normalized speaker space. The speech units from different speakers are mapped onto the normalized space in order to average out the inter-speaker variations as much as possible. The GMM-HMM models trained on the normalized acoustic features tend to have lesser mismatch if the test data is also normalized using the same approach. Hence, the normalization techniques are generally applied to both the training and the test data [77]. In the following we discuss some of the prominent feature-space normalization techniques.

2.5.1 Cepstral mean and variance normalization

Cepstral mean normalization (CMN) [78] and cepstral variance normalization (CVN) are very common and simple techniques to normalize the acoustic features derived from the raw speech. These transforms tend to sphere the data, i.e., each of the dimensions in the transformed feature has zero mean and unit variance. Given a set of D-dimensional cepstral features $\mathbf{O}$ such that its a series of L observation vectors $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_l, \ldots, \mathbf{o}_L\}$, the estimate for the d^{th} dimension of the l^{th} frame in the mean normalized feature is given as per the following relation [77]

$$\hat{o}_l(d) = o_l(d) - \mu(d) \tag{2.34}$$

where $\mu(d)$ is the d^{th} dimension in the mean of the observation vectors given by

$$\mu(d) = \frac{1}{L}\sum_{l=1}^{l} o_l(d).$$

Similarly, variance normalized feature is computed using the following equation

$$\hat{o}_l^{\text{CVN}}(d) = \frac{o_l(d) - \mu(d)}{\sqrt{\sigma(d,d)}} \tag{2.35}$$

where

$$\sigma(d,d) = \frac{1}{L}\sum_{l=1}^{L} \left(\hat{o}_l(d)\right)^2 .$$

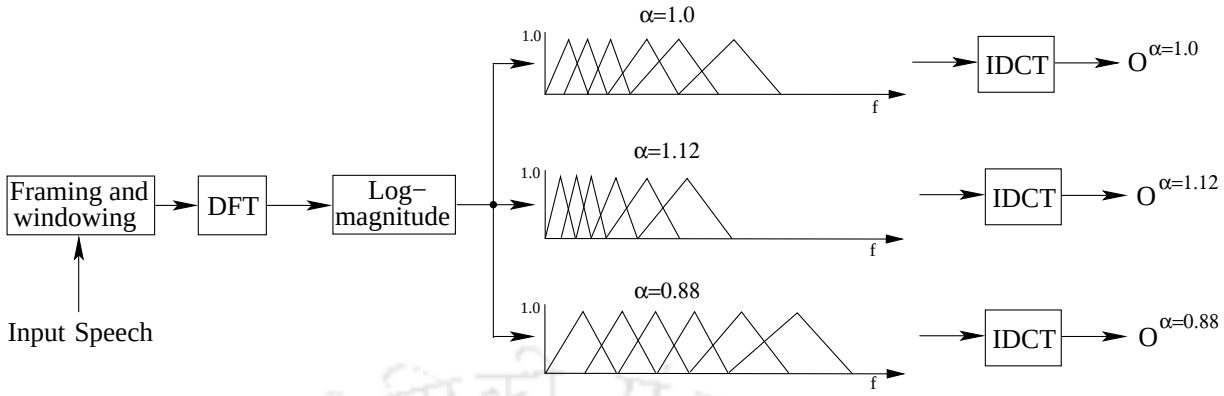

Figure 2.3: A block diagram illustrating the computation of VTLN-warped MFCC features through the compression/dilation of the triangular Mel-filterbank.

The advantage of this technique is that the normalization transform is estimated directly from the features with no transcription or model parameters being involved. Hence, this approach is computationally very simple. The main shortcoming of this technique is in the transform being global and therefore the ability to compensate for the acoustic mismatch is very limited [77]. Longer segments yield better estimates but result in more delay. Variable length silence in the speech utterance and very short utterances lead to degradation in performance. This technique also becomes infeasible for real-time applications since the entire segment must be procured before the normalization is performed.

2.5.2 Vocal tract length normalization

A major non-speech aspect affecting the performance of an ASR system is the variability in human voice among different speakers due to differences in their respective vocal tract lengths. The vocal tract lengths of male and female speakers are quite different (can vary from approximately 13 cm for females to over 18 cm for male speakers). As the vocal resonances are inversely proportional to the vocal tract length, the formant centre frequencies can vary by as much as 25% among speakers [26]. This leads to significant variation in the format structure for the same sound units across the speakers. Vocal tract length normalization (VTLN) [26] is used to reduce this mismatch among the speakers.

The basic idea behind VTLN is to define a warping factor α that compresses or expands the frequency domain during the feature extraction in order to map the actual speech signal to a normalized speech having lesser variability due to vocal tract length differences. To compress the speech signal in the frequency domain, the frequency scale of the speech signal is kept same

but the frequency scale of the Mel-filterbanks is stretched. Similarly, the filterbank frequencies are compressed to effectively stretch the signal in frequency domain [26]. VTLN requires the estimation of distinct warping factor for each speaker in the training as well as test data. A grid search is often used to obtain the optimal warping factor [79,80]. To do so, a set of discrete values of α_i are used as potential candidates. The value of α_i that maximizes the likelihood of the data is considered to be optimal, $\hat{\alpha}$ [26]. Let, $\mathbf{O}_s = \{\mathbf{o}_{s_1}, \mathbf{o}_{s_2}, \ldots, \mathbf{o}_{s_l}, \ldots, \mathbf{o}_{s_L}\}$ represent L frames of the observation sequence from a particular speaker s and let $\mathbf{\Lambda}$ denote the SI model. The optimal warping factor is then obtained as

$$\hat{\alpha}_s = \arg\max_{\alpha_i} P\big(\mathbf{O}_s^{\alpha_i} | \mathbf{\Lambda}\big) \tag{2.36}$$

where $\mathbf{O}_s^{\alpha_i}$ are the different frequency warped versions of the given observation sequence. Once the optimal warping factors are obtained for each speaker in the training set, the warped utterances are then used to develop a normalized GMM-HMM system $\mathbf{\Lambda}_N$. During decoding, the test observation sequence is warped to match the normalized model $\mathbf{\Lambda}_N$. Since, only the test utterance is available with no transcription, a three step process is used to estimate the corresponding optimal warping factor [26].

(i) $\mathbf{\Lambda}_N$ is used to decode the unwarped test sequence $\mathbf{O}_t$ to obtain the first-pass transcription.

(ii) The first-pass transcription is used to align each of the warped versions of the test data $\mathbf{O}_t^{\alpha_i}$ against the model $\mathbf{\Lambda}_N$ and the optimal warp factor $\hat{\alpha}$ is determined.

(iii) The test utterance is optimally warped using $\hat{\alpha}$ and then decoded using $\mathbf{\Lambda}_N$.

The outlined procedure happens to be computationally intensive. Hence, an alternate procedure to optimize the value of the warp factor is also proposed in [26]. For each α_i, the unwarped utterances that produce the highest likelihood for that value of warp factor are used to train GMMs to represent the feature-space distribution for that warp factor. During recognition, the likelihood of the test utterance is evaluated against each of the distribution and the warping factor corresponding to the distribution yielding highest likelihood is chosen as the optimal one. Though this approach is less complex, it suffers from the fact that it does not exploit the temporal information present in the speech signal which is modeled using HMMs.

2.5.3 Gaussianisation

Channel distortions tend to modify the speech feature distributions, e.g., the linear channel effects will shift the mean of the MFCC while the additive noise will tend to modify the variance.

Gaussianisation transforms the raw distorted features non-linearly so that the cumulative density function (CDF) of those features matches a predefined function, i.e., the standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. This makes the features more robust to channel and noise effects [81–83]. The CDF of the feature vector can be either parametric where histograms with discrete bins are applied or non-parametric having a CDF derived from Gaussians. Since the CDFs are defined for one-dimensional data only, the Gaussianisation must be performed separately for each of the dimensions [81].

2.6 Summary

In this chapter, we have presented a discussion on some of the dominant model-space adaptation and feature-space normalization techniques. The main focus of this chapter was on the review of the recently proposed acoustic model-interpolation-based fast adaptation techniques in the context of GMM-HMM-based ASR systems. We will be revisiting some of these techniques throughout the thesis. Furthermore, some of the discussed techniques will be combined with the approaches proposed in this work for enhancing the recognition performances. In addition to that, the feature-space normalization techniques, especially the VTLN, will be incorporated with all the approaches proposed in this thesis.

Indian Institute of Technology Guwahati

3

Assessing Fast Adaptation Approaches for Mismatched ASR

Contents

The work presented in this thesis is intended towards those applications that involve human-machine interactions as already mentioned. In the case of the interactive ASR tasks, adaptation needs to be performed either in the utterance-specific or in the incremental mode. Moreover, since no transcription of the adaptation data is available, the first-pass hypothesis of the adaptation data generated using the SI system should be used. But that being erroneous dilutes the effectiveness of the adaptation approaches significantly. In addition to that, the complexity of the employed adaptation technique is also a major issue in order to keep the overall latency low. The fast adaptation approaches discussed in Chapter 2 happen to be suitable for such tasks due to the low complexity involved in such techniques. Consequently, we have explored the effectiveness of the fast adaptation approaches in the context of interactive ASR tasks.

Apart from exploring some of the existing fast approaches, adaptation techniques that are found to have low computational cost are also presented. The main contributions of this chapter are two-fold: a reduced complexity dictionary-based basis search approach is presented and a novel basis selection criterion based on the minimization of the representation error is proposed. This is achieved by employing the greedy sparse representation techniques for the basis selection. A drawback of this is that SR is used only for the basis selection while the basis coefficients in sparse coding are discarded. Like the other model-interpolation-based techniques, the interpolation weights are estimated iteratively in the ML sense. The latency in the basis selection process is reduced due to the greedy SR-based approaches but the complexity of the weight estimation remains the same. Consequently, we have also presented a novel use of sparse coding to derive the adapted model mean parameters.

This chapter is organized as follows: In Section 3.1, the low complexity dictionary-based basis search approach is introduced. In Section 3.2, the improved basis selection scheme employing joint representation is discussed. The experimental setup used for the evaluation of the explored fast adaptation techniques is given in Section 3.3 while the evaluation results are presented in Section 3.3.4. In Section 3.4, we describe the proposed scheme that employs sparse coding to derive the adapted Gaussian mean parameters. We have also discussed our attempts to increase the degrees of freedom in this section. The evaluation of the proposed SR-based adaptation schemes are discussed in Section 3.4.4. Finally the chapter is summarized in Section 3.5.

3.1 Low Complexity Bases Search

As discussed earlier, the dynamic selection of basis is found to be more effective in the case of the model-interpolation-based fast adaptation techniques. One of the most common technique used for basis selection is the ML search involving the Viterbi-based alignment [62]. But this approach becomes cumbersome when the set of speakers from which the bases are to be selected is large. Motivated by this, in the following, we present a dictionary-based approach to reduce the computational cost involved in the basis search.

In the ML basis search, the test data is aligned against each of the SA models under the constraint of the first-pass hypothesis generated using the SI model. Under IID assumption, the log-likelihood of an observation sequence $\mathbf{O} = (\mathbf{o}_1, \ldots, \mathbf{o}_l, \ldots \mathbf{o}_L)$, with respect to the s^{th} speaker model, $\mathbf{\Lambda}^s$, given the state sequence $\mathbf{q} = (q_1, \ldots, q_l, \ldots, q_L)$, is

$$\log\Big(P\big(\mathbf{O}, \mathbf{q}|\mathbf{\Lambda}^s\big)\Big) = \sum_{l=1}^{L} \log(a_{q_l,q_{l+1}}) + \sum_{l=1}^{L} \log\big(b_{q_l}(\mathbf{o}_l)\big) \tag{3.1}$$

where,

$$b_{q_l}(\mathbf{o}_l) = \sum_{r=1}^{R} \left[\frac{w_{q_l}^{r,s}}{\sqrt{(2\pi)^n \mid \bar{\mathbf{C}}_{q_l}^r \mid}} \exp\left\{ -\frac{1}{2} (\mathbf{o}_l - \boldsymbol{\phi}_{q_l}^{r,s})^T (\bar{\mathbf{C}}_{q_l}^r)^{-1} (\mathbf{o}_l - \boldsymbol{\phi}_{q_l}^{r,s}) \right\} \right] \tag{3.2}$$

with $a_{q_l,q_{l+1}}$, $w_{q_l}^r$ and $\bar{\mathbf{C}}_{q_l}^r$ denoting the transition probability from state q_l to q_{l+1}, the Gaussian mixture-weights and the covariance matrix for r^{th} mixture component of state q_l in the SI model, respectively. The mean parameter for the r^{th} Gaussain in the s^{th} speaker model is denoted by $\boldsymbol{\phi}_{q_l}^{r,s}$.

From (3.1), note that in the computation of the likelihood with respect to the different SA models, the only term that varies is $\boldsymbol{\phi}_{q_l}^{r,s}$. Exploiting this fact, we devised a dictionary-based scheme for the selection of the bases. First, the SA models are derived for each of the training speakers using mean-only MAP adaptation of the SI model. From each of these SA models, a corresponding supervector representation (referred to as the *atom*) is derived by the concatenation of the mean parameters of all the Gaussian components in the corresponding SA model. All the atoms are collected in the form of an matrix referred to as the *exemplar dictionary* which is used for the basis selection. Similarly, a *target* is derived for each of the test utterances/speakers by performing mean only MAP adaptation under the constraints of the first-pass hypotheses. If $\boldsymbol{\phi}^s$ and $\mathbf{x}$ represent the atom and the target, respectively, the similarity between them can be computed by finding the correlation as $\frac{\mathbf{x}^T \boldsymbol{\phi}^s}{\|\mathbf{x}\| \|\boldsymbol{\phi}^s\|}$.

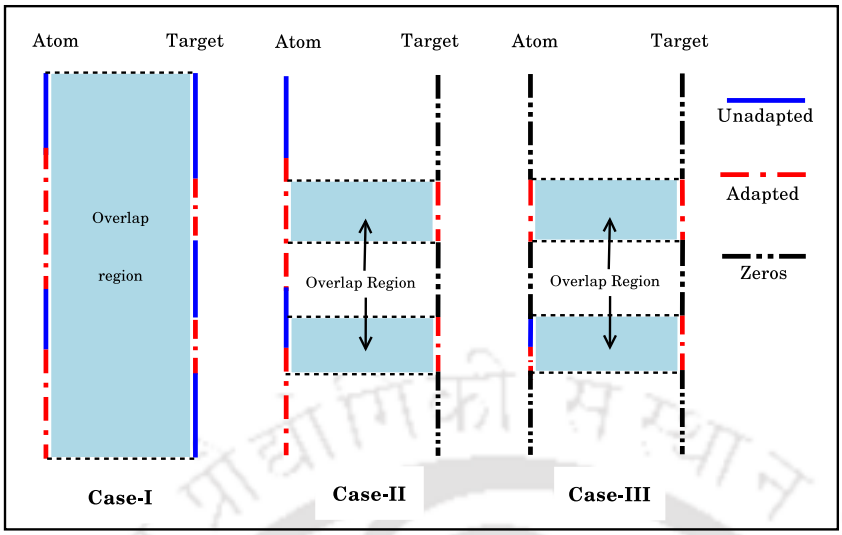

Figure 3.1: Illustrations of proposed conditionings for achieving the correct similarity score between a target and the dictionary atoms for basis selection.

In the above discussed approach, the selection of bases over N speaker adapted models requires only *one* alignment with respect to the SI model for deriving $\mathbf{x}$. On the other hand, the ML search requires N Viterbi alignments with respect to each of the SA models. Hence, in the proposed approach, there is a saving of $(N-1)$ Viterbi alignments per utterance. Furthermore, in the proposed technique, in contrast to the ML search, the mixture-weights are neglected and the posterior-weighted observations are used in the place of the original observations. Despite these differences, the two approaches result in similar adaptation performances as discussed later in Section 3.4.4.

3.1.1 Issues in creation of dictionary and target signal

In the context of the low-data adaptation task explored in this work, just finding the correlation does not yield a correct measure of the acoustic similarity. This point is explained with the help of the illustrations shown in Figure 3.1.

Case I: The default case of finding the similarity between an atom in the dictionary and a target is illustrated. In this case, all the dimensions of the two supervectors contribute to the similarity score. On account of data being limited, only a small number of the dimensions in the target get adapted (about 5-10% of the Gaussians get adapted for the WJSCAM0 database [84]). Hence the contribution to the similarity score happens to be dominated by the unadapted dimensions. This case simply results in a poor selection of the bases.

Table 3.1: Recognition performances for the different approaches for supervised/unsupervised adaptation in the utterance-specific adaptation mode. The given WERs for the adults' matched case ASR.

Utterance specific adaptation		WER (%)
Baseline (unadapted SI)		11.30
Most likelihood SA model		11.50
RSW	Unsupervised bases search & unsupervised weight estimation	10.57
	Supervised bases search & unsupervised weight estimation	10.27
	Supervised bases search & supervised weight estimation	10.09
Supervised global MLLR		9.66

Case II: It illustrates the proposed conditioning of the target to address the above mentioned shortcoming. In this case, the unadapted dimensions in the target are set to zero. In this way only the adapted Gaussains in the target contribute to the similarity score. This is similar to the case of the Viterbi-alignment-based ML search. In that case too, only the seen Gaussians in the adaptation data contribute to the likelihood score.

Case III: It illustrates another possibility of conditioning the target and the atoms. In this case, those unadapted dimensions that are set to zero in the target are also set to zero for each of the atoms. In this case too, the similarity score will be contributed by the adapted Gaussians as in Case II.

3.2 Basis Selection using Joint Representation

3.2.1 Motivation

In typical training corpus used for developing ASR systems, the data for each of the speakers does not cover all the phonetic contexts. For WSJCAM0 corpus used in this work, only 65-85% triphones get adapted while creating the SA models. Since each of the SA models have an insufficient coverage for all phonetic contexts, the simple most-likely SA model based adaptation is not very effective. A linear interpolation of acoustic models for adaptation (e.g., improved reference speaker weighting referred to as the RSW [62] henceforth) appears to address this shortcoming. The same can be inferred from the results given in Table 3.1. Furthermore, the use of the first-pass hypothesis for basis selection also affects the adaptation performance

due to the errors present in it. To assess the same, we have explored the basis selection using the true transcription while still estimating the interpolation weights in an unsupervised manner for the utterance-specific mode of adaptation. From Table 3.1, on comparing the recognition performances for the basis selection with and without the true transcription, the impact of errors can be assessed. Further, to assess the affect of errors on the interpolation weight estimation, the basis selection as well as the interpolation weights estimation is done in supervised manner. The adaptation performance for this study is also shown in Table 3.1. To benchmark these performances and to highlight the sub-optimality of the model-interpolation-based adaptation approach, the performance for the true-transcription-based global MLLR is also given in Table 3.1.

In the case of ML- or the correlation-based search, each basis is selected independently with respect to the seen phonetic contexts in the adaptation data. The unseen phonetic contexts in the selected bases, on the other hand, are not taken into account. The interpolation weights are also estimated using the seen contexts only. Hence, the bases are generally assigned weights proportional to their likelihood with respect to the adaptation data. On interpolating such bases, there will be varying degree of boosting of the contexts in the synthesized model as illustrated in Figure 3.2. In other words, the modelling of the phonetic space of the synthesized model would turn out to be unbalanced. To address this, we hypothesize that the bases should be selected in an orthogonal fashion by jointly minimizing the representation error of the target as discussed in the following subsections. Before outlining the proposed adaptation technique, we first present a brief overview of the sparse representation algorithms used in this work for joint representation.

3.2.2 Review of the sparse representation approaches

Sparse representation is a technique to solve ill-posed system of linear equations. The SR problem can be mathematically stated as: given a target vector $\mathbf{x} \in \mathbb{R}^M$ (referred to as the *signal*) and an exemplar dictionary $\mathbf{\Phi} = [\boldsymbol{\phi}_1 \dots \boldsymbol{\phi}_N] \in \mathbb{R}^{M \times N}$ (the columns are referred to as the *atoms*), find a vector $\boldsymbol{\upsilon} \in \mathbb{R}^N$ such that

$$\min \|\boldsymbol{\upsilon}\|_0 \quad \text{subject to} \quad \|\mathbf{\Phi}\boldsymbol{\upsilon} - \mathbf{x}\|_2 < \epsilon \tag{3.3}$$

where $\epsilon \in \mathbb{R}$ and $\epsilon > 0$. Due to the ℓ_0-norm minimization, the coefficient vector $\boldsymbol{\upsilon}$ turns out to be sparse. Thus only some atoms out of N are involved in representing a target. There are a number of techniques proposed in literature to solve the SR problem. These techniques

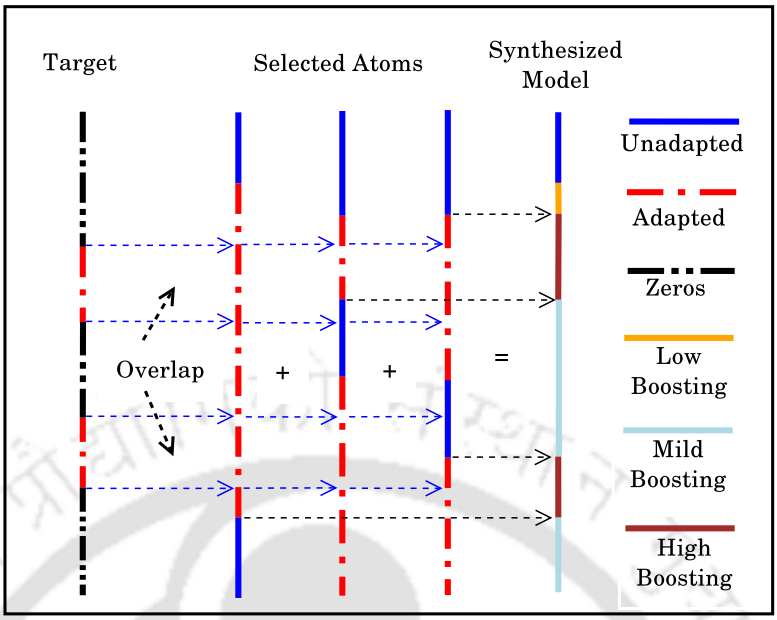

Figure 3.2: A schematic description for the interpolation of bases selected using the ML or the correlation search. The overlap region corresponds to those Gaussians in the test data that contribute to the likelihood/correlation score. In practical cases, some of the Gaussians in the atoms remain unadapted. Consequently, in the synthesized model, varying boosting of the phonetic contexts takes place depending on their relative states in the selected atoms and the interpolation weights. These weights are global in nature and estimated with respect to the seen Gaussians only. Hence those weights are not able to compensate for the different states of adaptation for all the phonetic contexts in the selected atoms.

can be classified into the family of *greedy* algorithms and *relaxation* algorithms. The greedy algorithms attempt to construct the support one atom at a time. On the other hand, the relaxation techniques smooth the ℓ_0-norm and solve the problem using the classical methods from continuous optimization. In the following subsections, we provide an overview of both the kinds of SR techniques.

3.2.2.1 Matching and orthogonal matching pursuits

The matching pursuit (MP) [68] and the orthogonal matching pursuit (OMP) [85] are greedy techniques intending to construct a sparse representation through an iterative process. A locally optimum solution is obtained at each iteration with an intention to obtain the global optimum at the end of the iterations. The MP/OMP takes only as many iterations as the desired sparsity to approximate the solution and are consequently very fast. In the case of MP, an atom may be selected more than once making the matrix of selected atoms non-invertible. The OMP algorithm addresses the issues in the MP and is outlined in the following.

Let $\mathbf{\Phi}$ be the dictionary with N atoms and $\mathbf{x}$ be the signal to be approximated. To begin with, a residual vector $\boldsymbol{y}$ is initialized to be equal to the signal, i.e., $\boldsymbol{y}_0 = \mathbf{x}$. This is followed by finding the atom of $\mathbf{\Phi}$ that is most correlated to the residual vector as

$$\omega_1 = \arg\max_i |< \boldsymbol{y}_0, \boldsymbol{\phi}_i >|, \quad i = 1, \ldots, N. \tag{3.4}$$

In (3.4), the index of the most correlated atom is given by ω_1. Solving a least square-problem the vector $\boldsymbol{\upsilon}_1$ is derived as given

$$\arg\min_{\boldsymbol{\upsilon}_1} \|\mathbf{x} - \tilde{\mathbf{A}}_1 \boldsymbol{\upsilon}_1\|_2 \tag{3.5}$$

where $\tilde{\mathbf{A}}_1$ is the matrix of selected bases such that

$$\tilde{\mathbf{A}}_1 = [\boldsymbol{\phi}_{\omega_1}].$$

The residual vector is then updated as follows:

$$\boldsymbol{y}_2 = \mathbf{x} - \tilde{\mathbf{A}}_1 \boldsymbol{\upsilon}_1. \tag{3.6}$$

This process is iterated a number of times until the representation error becomes less than ϵ or the desired sparsity value has been achieved. For any particular iteration k, following steps are involved:

(i) Determine the most correlated atom with respect to the current residual

$$\omega_k = \arg\max_i |< \boldsymbol{y}_{k-1}, \boldsymbol{\phi}_i >|, \quad i = 1, \ldots, N$$

(ii) The selected atom is added to the support

$$\tilde{\mathbf{A}}_k = \tilde{\mathbf{A}}_{k-1} \cup \boldsymbol{\phi}_{\omega_k} \quad \text{where} \quad \tilde{\mathbf{A}}_{k-1} = [\boldsymbol{\phi}_{\omega_1} \ \ldots \ \boldsymbol{\phi}_{\omega_{k-1}}]$$

(iii) Find $\boldsymbol{\upsilon}_k$

$$\arg\min_{\boldsymbol{\upsilon}_k} \|\mathbf{x} - \tilde{\mathbf{A}}_k \boldsymbol{\upsilon}_k\|_2$$

(iv) Determine the updated residual vector

$$\boldsymbol{y}_k = \mathbf{x} - \tilde{\mathbf{A}}_k \boldsymbol{\upsilon}_k$$

(v) Stop if

$$\|\boldsymbol{y}_k\|_2 < \epsilon$$

or if the desired sparsity (K) is attained

After each iteration, the atom that best describes the residual is added to the support along with the basis coefficient in sparse coding that contributes the most in reducing the residual. In order to make sure that a column of $\boldsymbol{\Phi}$ is selected only once, unlike the MP, the basis coefficients in sparse coding are re-computed for the entire set of support each time an atom is added. Consequently, the currently added atom is orthogonal to the set of previously selected atoms and hence the term 'orthogonal' appears in the name of algorithm.

3.2.2.2 Least absolute shrinkage and selection operator

In the least absolute shrinkage and selection operator (Lasso), instead of minimizing the ℓ_0-norm, the solution is optimized by minimizing the representation error under the constraint of given sparsity. At the same time, the ℓ_0-norm is replaced by the ℓ_1-norm making the problem easier to work with. In the case of the Lasso, the problem is formulated as

$$\|\boldsymbol{\Phi}\boldsymbol{\upsilon} - \mathbf{x}\|_2 < \epsilon \qquad \text{subject to} \qquad \|\boldsymbol{\upsilon}\|_1 < K$$

where K is the desired sparsity. The Lasso can hence be viewed as an optimization problem and there are a number of convex optimization techniques available to solve such problems involving ℓ_1-norm. Therefore, it can be either solved as a linear programming problem when $\epsilon = 0$ or a quadratic programming when $\epsilon \neq 0$. The simplex and the interior-point [86] methods are some of the approaches that are reported for such optimization problems. Least angle regression (LARS) [87] is a greedy approach to the Lasso and has comparatively much lesser complexity than the two former techniques.

3.2.3 Proposed SR-based basis selection approach

The joint representation-based basis selection problem can be interpreted as finding a sparse representation for a given target over an exemplar dictionary. As already mentioned, there are a number of techniques proposed in literature to solve this problem. In this work, OMP and LARS algorithms are used for SR due to their lower complexity [64]. The proposed fast adaptation approach using OMP-based SR is outlined in Algorithm 2. We have chosen OMP because of its least complexity among all the other SR techniques.

Table 3.2: Specifications of the speech corpora used for the experimental evaluation presented in this work.

Corpus	WSJCAM0		PF-STAR	
Language	British English		British English	
Data set	**CAMtr**	**CAMts**	**PFtr**	**PFts**
Purpose	Training	Testing	Training	Testing
No. of speakers	92	20	122	60
Speaker kind	Adults (male/female)	Adults (male/female)	Children (male/female)	Children (male/female)
Age group	$>$ 18 years	$>$ 18 years	4-15 years	4-15 years
No. of words	132,778	5,608	46,974	5,067
Duration (hrs.)	15.5	0.6	8.3	1.1
Recording conditions	Quite room, close talking and desk microphones		Closed room, head mounted microphones	

Unlike the ML search, in the proposed approach each basis is selected with respect to a residual instead of the target itself (see Step 6). This residual is updated as the error between the target and its current estimate using the selected bases (see Step 9). After the selection of the first basis, on the computation of the residual, the dimensions that were set to zero in the target also become active. These are now initialized with the values derived from the most correlated selected atom. Thus, in successive selections, even the unadapted dimensions in the target start contributing in the basis selection. Moreover, each new selected basis adds to the spanned phonetic space in an orthogonal manner. Consequently, the adapted model synthesized by the selected bases ensures a balanced coverage of all phonetic contexts. This is expected to remove the errors in the first-pass hypothesis in a consistent manner.

3.3 Experimental Evaluation

In the following subsections, the details of the setup used in the experimental evaluations performed in this thesis are discussed.

Algorithm 2 Proposed joint-representation-based basis selection for fast adaptation

Given: The exemplar dictionary $\mathbf{\Phi}$ obtained by collecting the speaker-specific mean-supervectors as the columns of the matrix, $\mathbf{\Phi} = [\boldsymbol{\phi}_1 \ldots, \boldsymbol{\phi}_j, \ldots, \boldsymbol{\phi}_N]$, the number of bases to be selected K, the SI model parameters $\mathbf{\Lambda}$

Step 1: Obtain the first-pass hypothesis for test data by decoding with $\mathbf{\Lambda}$

Step 2: MAP adapt the SI mean vectors using the given test data using the first-pass hypothesis with relevance factor (τ) set to zero

Step 3: To form the target $\mathbf{x}$, extract the mean-supervector and substitute zeros for all unadapted dimensions

Step 4: Initialize the residual $\boldsymbol{y}_0 = \mathbf{x}$ and the matrix of the selected bases $\tilde{\mathbf{A}}_0 = \phi$

for $k = 1,\ k \leq K,$ **do**

Step 5: Find the index of the most correlated atom

$$\omega_k = \arg\max_i |< \boldsymbol{y}_{k-1}, \boldsymbol{\phi}_j >|, \quad i = 1, \ldots, N$$

Step 6: Update the matrix of the selected bases

$$\tilde{\mathbf{A}}_k = \tilde{\mathbf{A}}_{k-1} \cup \boldsymbol{\phi}_{\omega_k} \quad \text{where} \quad \tilde{\mathbf{A}}_{k-1} = [\boldsymbol{\phi}_{\omega_1} \ \ldots \ \boldsymbol{\phi}_{\omega_{k-1}}]$$

Step 7: Minimize the representation error for $\mathbf{x}$ by solving the least-square problem

$$\arg\min_{\boldsymbol{\upsilon}_k} \|\mathbf{x} - \tilde{\mathbf{A}}_k \boldsymbol{\upsilon}_k\|_2$$

where $\boldsymbol{\upsilon}_k$ is the vector of basis coefficients in sparse coding after k^{th} iteration

Step 8: Update the residual

$$\boldsymbol{y}_k = \mathbf{x} - \tilde{\mathbf{A}}_k \boldsymbol{\upsilon}_k$$

end for

Step 9: For model interpolation, choose the columns from the exemplar dictionary $\mathbf{\Phi}$ corresponding to the indices of the selected bases in $\tilde{\mathbf{A}}_K$

Step 10: Estimate the interpolation weights $\boldsymbol{\eta} = [\eta_1 \ldots \eta_K]^T$ in the ML sense following the iterative procedure given in Algorithm 1

Step 11: The estimate of the mean supervector for the adapted model is given as

$$\hat{\boldsymbol{\phi}} = [\boldsymbol{\phi}_{\omega_1} \ \ldots \ \boldsymbol{\phi}_{\omega_K}]\boldsymbol{\eta}$$

Step 12: The adapted GMM-HMM model is obtained by replacing the mean vectors in $\mathbf{\Lambda}$ by their corresponding values in $\hat{\boldsymbol{\phi}}$

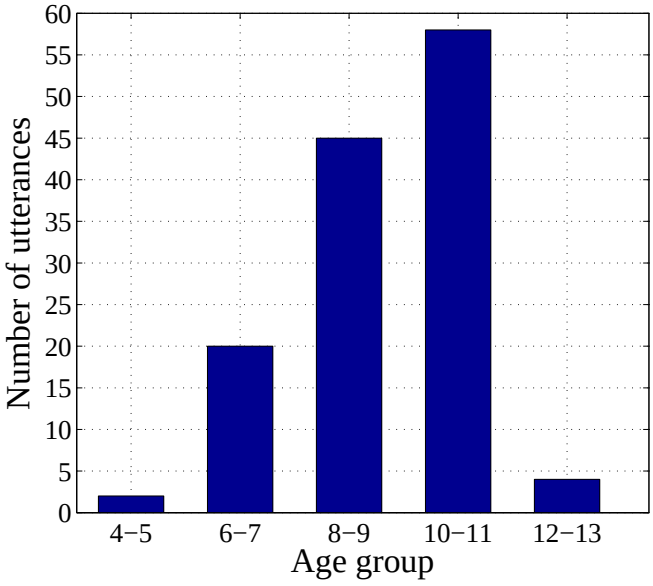

Figure 3.3: A bar graph depicting the number of utterances in PFts belong to a particular age group.

3.3.1 Speech corpus

For the continuous speech recognition task, an ASR system is developed using the WSJ-CAM0 Cambridge Read News speech corpus [84]. This database contains speech data from 92 adult male/female speakers for training. The training set (CAMtr) consists of 7861 utterances with approximately 90 sentences per speaker. The total duration of training data is 15.5 hours. In order to evaluate the matched case recognition performance of the developed ASR system, the CAMts test set was used. This test set consists 0.6 hours of data from 20 adult speakers constituting a total of 5608 words. All speech data is re-sampled to 8 kHz rate [1]. More details about the WSJCAM0 speech corpus are summarized in Table 3.2. For evaluating the effectiveness of the basis search schemes in those cases where there happens to be a greater acoustic mismatch, decoding of children's speech data is performed. The PF-STAR speech corpus [88] is used for testing the effectiveness of the explored approaches for the recognition of children's speech on adults' speech (CAMtr) trained acoustic models. The test set of PF-STAR (PFts) contains 1.1 hours (129 utterances) of speech data from 60 children. The total of words in this test set is 5067. More details about this speech corpus are also given in Table 3.2. The age-wise split of the number test utterances in PFts is shown in Figure 3.3.

[1] In the end of this thesis, we repeat some of the experiments on speech data sampled at 16 kHz rate to re-validate the salient inferences.

3.3.2 ASR system specifications

The recognition performances are evaluated on the ASR system developed using the HTK toolkit [9]. A Hamming window of length 25 ms keeping a frame rate of 100 Hz is used for short-time analysis of the speech data. A pre-emphasis factor of 0.97 is used. Employing a 21-channel Mel-filterbank, first 12-dimensional base MFCC features are computed (C_1-C_{12})[2]. The log power is added as the zeroth coefficient making the base feature dimension equal to 13 (C_0-C_{12}). These base features are then augmented with their first- and second-order temporal derivatives (computed over a span of 5 frames) to yield the 39-dimensional features used in the acoustic modelling. The cepstral mean and variance normalization are also applied to all features. Throughout this chapter, the so computed 39 dimensional features are referred to as the *default* features.

The baseline SI system is developed using cross-word triphone acoustic model training along with decision-tree-based state tying. The GMM-HMM-based approach is employed for statistical modelling. Each triphone model consists of a 3-states HMM with 8 diagonal covariance Gaussian components per state. A 3-state HMM with 16 diagonal covariance Gaussian components per state is used for modelling the silence and the short-pause. In the matched case, the developed ASR system performance is evaluated using the CAMts test set. The standard WSJ0 5,000 words closed non-verbalized vocabulary set and the MIT-Lincoln 5k Wall Street Journal bigram language model (LM) are employed in the decoding of the CAMts test. The used LM has a perplexity of 95.3 for the CAMts set with no out-of-vocabulary (OOV) words.

The word error rate (WER) is used as a measure of the recognition performances of the different techniques explored in this thesis. The WER is computed as follows:

$$\%\text{WER} = \frac{\text{SUB} + \text{DEL} + \text{INS}}{\text{Total number of words in the reference}} \times 100 \tag{3.7}$$

where, SUB, DEL and INS represent the number of substitutions, deletions and insertions made in the hypothesized text transcript with respect to the true transcription, respectively. The recognition performance of the developed baseline ASR system (unadapted SI system) turns out to be the same as that reported in [84].

3.3.3 Adaptation experiments

For the model-space-based fast adaptation approaches explored in this work, experiments are performed under the constraints of the first-pass hypothesis in the following two modes:

[2]Refer to Appendix A for a brief overview.

- Each of the test utterances is treated independently, i.e., the statistics required for adapted parameter estimation are computed for each test utterance separately (*utterance-specific*).

- For every speaker, the adaptation data is made available to the system in an incremental manner. In other words, when the n^{th} test utterance is made available, the statistics of all previous $(n-1)$ utterances are accumulated with the current statistics (*incremental*).

For the SR-based basis selection techniques, OMP is implemented using the OMP-Box v- 10 [89] while the SpasSM toolbox [90] is used for the Lasso implementation employing the modified LARS [87, 91]. For all the experiments, the number of acoustic models interpolated is varied from 4 to 18 and the WERs for the same are presented. As the number of bases increases, the computational cost of interpolation weight estimation increases, and hence only up to 18 bases are explored.

3.3.4 Results and discussion

The SA models employed in the model-interpolation-based fast adaptation may be derived using either the MAP or the MLLR transformation of the SI mean vectors. In our experiment, it was noted that the WERs for the cases when the MAP or the MLLR is employed for creating the SA model turned out to be quite similar. Hence, the MAP is employed for deriving the SA models in all the experimental evaluations reported in this thesis. The number of bases interpolated to derive the adapted Gaussian vectors is varied from 4 to 18 as mentioned earlier.

In Figure 3.4, the recognition performances are shown for the different basis selection schemes discussed in this chapter. Note that the low complexity correlation-based bases search results in a performance quite similar to that of the conventional ML search. The performances of the proposed joint-representation-based basis selection criterion are also given in Figure 3.4. Between the two conditioning considered, Case II resulted in better performance due to the somewhat balanced coverage of spanned phonetic space. On the other hand, for Case III conditioning, the same could not be achieved. This is so because the dimensions over which the correlation is computed is always restricted to those of the target. Furthermore, the best case performance of the proposed technique turns out to be 10.16% which is very close to that for the oracle experiment (10.09%) given in Table 3.1. The WERs to compare the OMP- and the LARS-based basis search in the two modes of adaptation are shown in Figure 3.5. Sim-

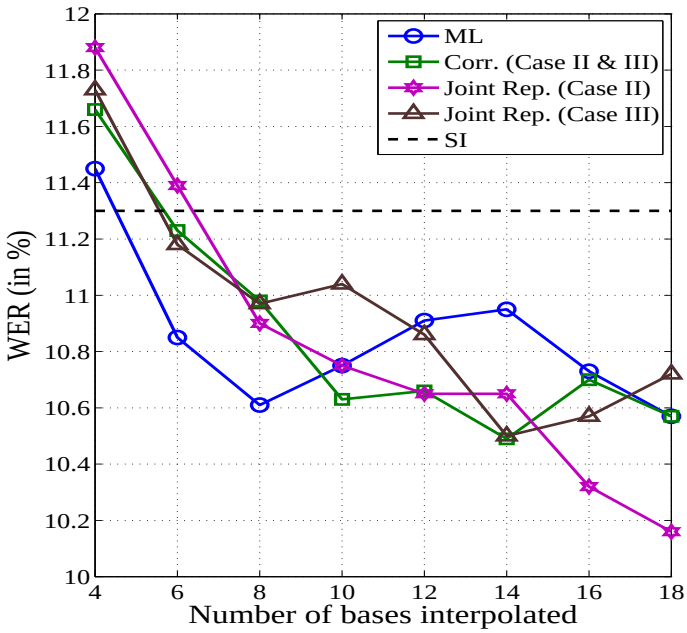

Figure 3.4: The utterance-specific WERs for basis selection using the ML search, the low complexity correlation search and the proposed joint representation for the two cases of conditioning.

ilar trends are noted for both the modes of adaptation. Moreover, there is a saturation in recognition performance with the increase adaptation data. This is attributed to the stringent constraints imposed on the degrees of freedom in the formulation of the model-interpolation-based approaches itself [54].

3.3.5 Experiments on highly mismatched (children) test set

As mentioned earlier, the mismatched testing is performed using the PFts test set from the PF-STAR speech corpus. There are significant differences in the word-list and the word frequencies across the adults' and the children's datasets. As a result, for achieving a meaningful recognition performance for the children's test set, a 1.5k bigram LM is employed. This domain-specific LM is trained on the transcripts of speech data in PF-STAR excluding PFts i.e., on the training transcripts only. In mismatched decoding, a lexicon of 1,969 words including pronunciation variations is used. The employed domain-specific LM has OOV rate of 1.02% and perplexity of 95.8 for the PFts set. This helps in overcoming the significant linguistic mismatches between the train and the test sets. The effect of using out-of-domain LM can be understood from the WERs given in Table 3.3

For the adaptation experiments, the number of bases interpolated is varied as mentioned earlier. Table 3.4 enlists the best case recognition performances for the various basis search

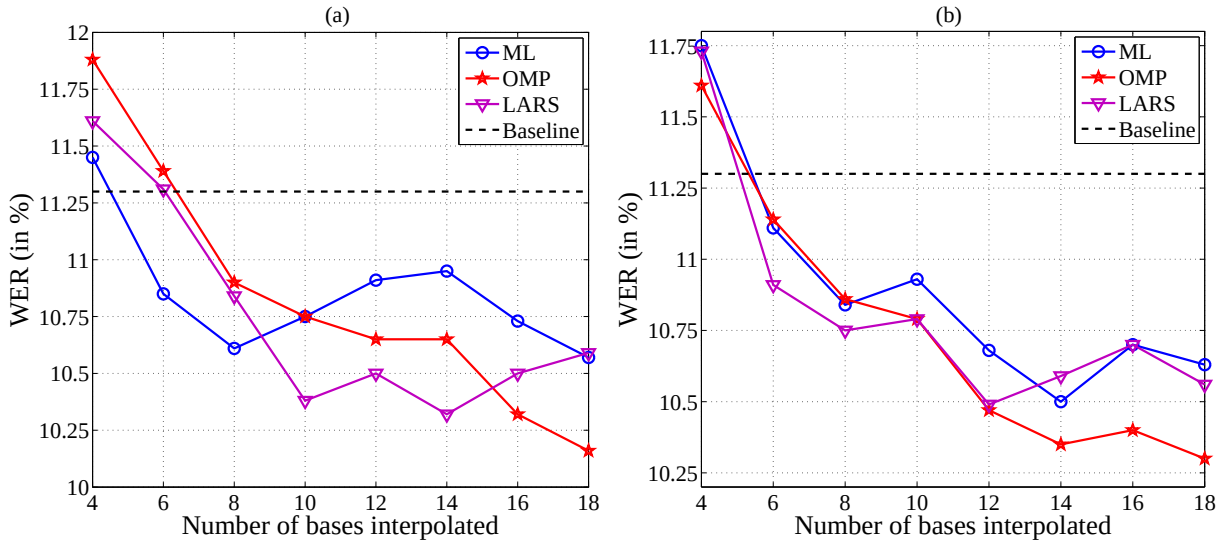

Figure 3.5: The recognition performances for basis selection using ML and the greedy SR techniques (OMP and LARS). The WERs are plotted in (a) utterance-specific and (b) incremental modes.

Table 3.3: Performances for the adults' speech trained SI system for the two test sets with and without using domain-specific LM. Also given are the out-of-vocabulary word rates and perplexities of the two LMs with respect to the two test sets.

LM domain	CAMts			PFts		
	Perpelxity	OOV	WER	Perpelxity	OOV	WER
Adult	95.3	0.00	**11.30**	86.4	19.53	110.97
Children	23.20	40.71	91.82	95.8	1.20	**64.24**

approaches using two different test sets. In this table, Corr-I corresponds to the case when unadapted dimensions are not replaced by zeros while the modified target case is denoted by Corr-II. It is evident that, with the suggested modifications, the correlation-based search leads to performances similar to that of the ML basis search. The recognition performances for the SR-based basis selection approaches are also given in Table 3.4. This scheme is found to outperform the ML search when evaluated on the two different test sets. The differences are more pronounced in the case of PFts due to increased acoustic mismatch. Though both the OMP and the LARS result in similar recognition performances, the latter has a greater computational cost. Therefore, the OMP has been employed for the SR in the remaining of this work.

Table 3.4: WERs for the ML-, the correlation- and the SR-based basis search approaches. The performances are evaluated using two different test sets, namely CAMts and PFts. The recognition performances are shown for the incremental (Incr.) and the utterance-specific (Utt.) modes of fast adaptation. The number of bases interpolated is varied and the WERs for 18 bases case are tabulated.

Basis Search Technique	WER (in %)			
	CAMts		PFts	
	Incr.	Utt.	Incr.	Utt.
ML	10.50	10.57	50.26	49.93
Cor- I	10.88	10.90	53.15	53.86
Corr-II	10.49	10.54	49.98	50.18
SR (OMP)	10.28	10.16	47.29	47.82
SR (LARS)	10.39	10.32	46.93	48.00
Unadapted SI	11.30		64.24	

3.4 SR-based Acoustic Model Adaptation

In the approach outlined in Section 3.2, the model interpolation is done using weights derived by an iterative ML estimation procedure. The work reported in [53] employed 4 iterations during weight estimation. But in our experiments, it is noted that 6 to 7 iterations are required to ensure the convergence. This accounts for the major portion of computational cost in the model-interpolation-based adaptation process. For reducing this computational cost, we explored the possibility of using the sparse coded target as the Gaussian mean parameter in the adapted model. Given the estimate of the sparse coefficient vector $\hat{\boldsymbol{v}}$ in the SR-based basis search, a sparse coded target vector can be determined as

$$\tilde{\mathbf{x}} = \boldsymbol{\Phi}\hat{\boldsymbol{v}}. \tag{3.8}$$

As the sparse coding process is fast, this approach is expected to result in substantial reduction in the cost of weight estimation when the number of bases to be interpolated is large. Unfortunately, our initial attempts did not succeed.

On analyzing, it was noted that the dictionary atoms and the target are required to be normalized to unit-norm in the sparse coding while no such constraint is imposed in the ML estimation of basis interpolation weights. Consequently, the coefficients for the bases in SR turn out to have different dynamic ranges than those of the corresponding ML weights. Therefore,

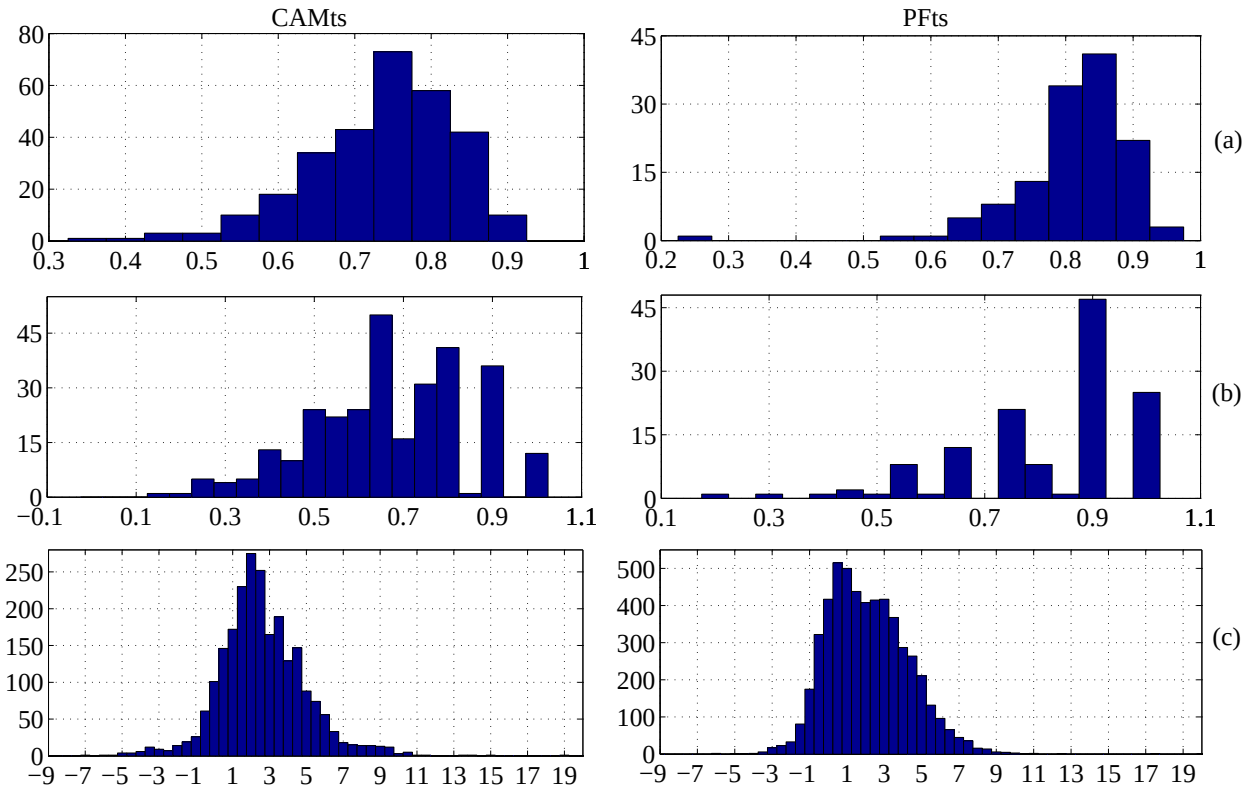

Figure 3.6: Studying the relation between the basis coefficients in sparse coding and the corresponding ML weights for all the utterances in CAMts and PFts test sets. Shown are the histograms of correlation coefficients between the two sets of vectors for (a) the magnitude only case, (b) the sign only case, and (c) the histogram of basis-wise ratio of ML weight to the corresponding basis coefficient for both test sets.

the sparse coded target cannot be used as the Gaussian mean parameter in the adapted acoustic model. This is so because its dynamic range differs from those of the Gaussian mean and the covariance parameters learned during the SI model training. Moreover, the sparse coding does not guarantee that the derived model means will enhance the likelihood of the adaptation data over the unadapted SI system. Enhanced likelihood, in turn, is expected to improve the recognition performance.

In order to address this mismatch in the dynamic ranges, we compared the basis coefficients in the SR and the estimated ML interpolation weights for the adaptation experiments performed with respect to the SI system. To do the same, the correlation between the two when 18 bases are interpolated to derive the adapted model, are studied. The correlation coefficients are computed for the magnitude as well as the sign of the basis coefficients in the SR and the corresponding ML basis weights, and their histograms are shown in Figures 3.6(a) and 3.6(b), respectively. It is evident that there exists a strong correlation between the two. In other

words, for almost all the test utterances, when the basis coefficients in the sparse coding are positive, the corresponding ML weights are also positive and vice-versa. Furthermore, if the bases are arranged in the order of the magnitude of the sparse coefficients and the ML weights, almost same ordering is observed in both the cases. Figure 3.6(c) shows the histogram of the basis-wise ratio of the ML weight to the corresponding basis coefficient for both the test sets. Interestingly, these distributions have turned out to be predominantly unimodal with a small spread. These observations motivated us to explore a single-point scaling of the sparse coded target such that its likelihood with respect to the SI model gets enhanced. In the following we present the approach for estimating a global scaling parameter using the ML criterion.

3.4.1 Estimation of maximum-likelihood scaling factor

Given the sparse coded target $\tilde{\mathbf{x}}$, the adapted Gaussian mean supervector $\boldsymbol{\phi}$ is modeled as

$$\boldsymbol{\phi} = \eta\tilde{\mathbf{x}} \tag{3.9}$$

where η is the global scaling factor to be estimated in the ML sense. As the scaling is global, we can separate the adaptation problem for each of R Gaussians in the SI model. Thus, the r^{th} Gaussian mean parameter vector in the adapted model can be written as

$$\boldsymbol{\phi}^r = \eta\tilde{\mathbf{x}}^r \tag{3.10}$$

where $r = 1, \ldots, R$ (the number of Gaussians in the GMM-HMM). Let, $\mathbf{O}$ be the adaptation data which is a series of L observation sequences, i.e., $\mathbf{O} = (\mathbf{o}_1, \ldots, \mathbf{o}_L)$; $\gamma^r(l)$ be the *a posteriori* probability of occupying the r^{th} Gaussian given that the observation sequence $\mathbf{o}_l$ is generated; $\bar{\mathbf{C}}^r$ be the covariance matrix for the r^{th} Gaussian component in the SI model and D be the dimension of the feature vector. Further, let the current and the updated HMM parameters be denoted by $\boldsymbol{\Lambda}$ and $\bar{\boldsymbol{\Lambda}}$, respectively. The *auxiliary function* $Q(\boldsymbol{\Lambda}, \bar{\boldsymbol{\Lambda}})$ [92] to be maximized for estimating η (treating the terms independent of $\boldsymbol{\phi}$ as constants) is given as

$$Q(\boldsymbol{\Lambda}, \bar{\boldsymbol{\Lambda}}) = -\frac{1}{2}P(\mathbf{O}|\boldsymbol{\Lambda})\sum_{p=1}^{P}\sum_{l=1}^{L}\gamma^r(l)b(\mathbf{o}_l, r) + \text{ const.} \tag{3.11}$$

where

$$b(\mathbf{o}_l, r) = \left(\mathbf{o}_l - \boldsymbol{\phi}^r\right)^T(\bar{\mathbf{C}}^r)^{-1}\left(\mathbf{o}_l - \boldsymbol{\phi}^r\right) \tag{3.12}$$

or

$$b(\mathbf{o}_l, r) = \left(\mathbf{o}_l - \eta\tilde{\mathbf{x}}^r\right)^T(\bar{\mathbf{C}}^r)^{-1}\left(\mathbf{o}_l - \eta\tilde{\mathbf{x}}^r\right). \tag{3.13}$$

On differentiating (3.11) with respect to η and setting it to zero, we have

$$-\frac{1}{2}P(\mathbf{O}|\mathbf{\Lambda})\frac{d}{d\eta}\sum_{r=1}^{R}\sum_{l=1}^{L}\gamma^r(l)\big[(\mathbf{o}_l-\eta\tilde{\mathbf{x}}^r\big)^T(\bar{\mathbf{C}}^r)^{-1}\big(\mathbf{o}_l-\eta\tilde{\mathbf{x}}^r)\big]=0 \tag{3.14}$$

On simplifying (3.14), the estimate of the scale factor can be expressed as

$$\hat{\eta}=\Bigg[\sum_{r=1}^{R}\Big(\sum_{l=1}^{L}\gamma^r(l)\Big)(\tilde{\mathbf{x}}^r)^T(\bar{\mathbf{C}}^r)^{-1}\tilde{\mathbf{x}}^r\Bigg]^{-1}\Bigg[\sum_{r=1}^{R}(\tilde{\mathbf{x}}^r)^T(\bar{\mathbf{C}}^r)^{-1}\Big(\sum_{l=1}^{L}\gamma^r(l)\,\mathbf{o}_l\Big)\Bigg] \tag{3.15}$$

The estimate of the r^{th} adapted mean vector is then given as

$$\hat{\boldsymbol{\phi}}^r=\hat{\eta}\tilde{\mathbf{x}}^r \tag{3.16}$$

Note that above estimation process is identical to that used in the cluster adaptive training except the weight being scalar. From (3.15), it is obvious that the computation of η is dependent on the computation of $\gamma^r(l)$. Therefore, η can be derived using an iterative approach similar to that outlined in the CAT scheme. In the computation of η, the initial value of $\gamma^r(l)$ is important for the convergence. In the following, we discuss the problem that arises in the estimation of the scaling factor due to improper initialization.

3.4.1.1 Issues in weight initialization

In the EV-based fast adaptation, the eigenvectors are derived using PCA performed on the correlation matrix created using the mean supervectors. Hence their linear combination is not guaranteed to result in a data likelihood higher than that obtained with respect to the SI model. To deal with this, in the EV approach, the adapted Gaussian mean parameter is modeled using a bias $\bar{\boldsymbol{\phi}}$ with an unity weight as follows

$$\boldsymbol{\phi}=\bar{\boldsymbol{\phi}}+\sum_{i=1}^{K}\eta_i\mathbf{e}_i \tag{3.17}$$

where $\{\mathbf{e}_i\}_{i=1}^{K}$ represents the K eigenvoices and $\{\eta_i\}$ is the set of corresponding interpolation weights. As the interpolation weight estimation is interlinked with the computation of $\gamma^r(l)$, so an improper initialization may lead to a poor estimate of the weights. The bias $\bar{\boldsymbol{\phi}}$, which is often chosen as the SI mean supervector, ensures that the initial value of $\gamma^r(l)$ is computed pivoted to the space of the SI model parameters. This, in turn, guarantees an increase in the likelihood of the adaptation data beyond that achieved by the SI model. To confirm this point, we performed adaptation experiments without the bias term in weight estimation.

In the case of the sparse coding, the bases are derived by minimizing the representation error. Therefore, like the eigenvectors, an improper initialization of $\gamma^r(l)$ leads to a degradation in the recognition performance. Employing the same reasoning, the adapted mean parameter for the r^{th} Gaussian is now modeled as

$$\boldsymbol{\phi}^r = \bar{\boldsymbol{\phi}}^r + \eta\tilde{\mathbf{x}}^r \tag{3.18}$$

where $\bar{\boldsymbol{\phi}}^r$ is the mean parameter vector of r^{th} Gaussian in the SI model. Using (3.18) in (3.12) yields

$$b(\mathbf{o}_l, r) = \big(\mathbf{o}_l - \bar{\boldsymbol{\phi}}^r - \eta\tilde{\mathbf{x}}^r\big)^T(\bar{\mathbf{C}}^r)^{-1}\big(\mathbf{o}_l - \bar{\boldsymbol{\phi}}^r - \eta\tilde{\mathbf{x}}^r\big). \tag{3.19}$$

With this modification, the estimate of η is given by

$$\hat{\eta} = \Bigg[\sum_{r=1}^{R}\Big(\sum_{l=1}^{L}\gamma^r(l)\Big)(\tilde{\mathbf{x}}^r)^T(\bar{\mathbf{C}}^r)^{-1}\tilde{\mathbf{x}}^r\Bigg]^{-1}\Bigg[\sum_{r=1}^{R}(\tilde{\mathbf{x}}^r)^T(\bar{\mathbf{C}}^r)^{-1}\Big(\sum_{l=1}^{L}\gamma^r(l)[\mathbf{o}_l - \bar{\boldsymbol{\phi}}^r]\Big)\Bigg] \quad . \tag{3.20}$$

Finally, the estimate of the r^{th} adapted mean vector is determined as

$$\hat{\boldsymbol{\phi}}^r = \bar{\boldsymbol{\phi}}^r + \hat{\eta}\tilde{\mathbf{x}}^r. \tag{3.21}$$

The pseudo-code of iteratively finding the scaling factor η is as follows:

Step 1. Begin with the scaling factor set to zero, $\eta = 0$

Step 2. Find the current value of adapted mean $\boldsymbol{\phi}^r$ using (3.18)

Step 3. Compute the Gaussian posterior matrix $\boldsymbol{\gamma}$ for the adaptation data using the current value of adapted mean supervector $\boldsymbol{\phi}$

Step 4. Update the value of scaling factor $\hat{\eta}$ using (3.20)

Step 5. Repeat steps 2, 3 and 4 until convergence

3.4.2 Sparse coding over learned dictionary

In the previous section, the proposed scaling of sparse coded target involved an exemplar dictionary (the atoms correspond to speaker-specific mean supervectors). In this work, we have also explored the effectiveness of the proposed approach on a *learned dictionary*. The eigenvectors that are employed as bases in the EV approach are learned from the speaker-specific mean supervectors using PCA. Consequently, the set of eigenvectors can be considered as a learned dictionary. Therefore, a similar study is also performed to evaluate the effectiveness

of the proposed approach in the eigenvector domain. The scheme outlined in [54] is followed to derive the eigenvectors. But, unlike the EV approach, all eigenvectors are retained to form a learned dictionary. The target is then sparse coded over that dictionary and the resulting sparse coded target is scaled to derive the adapted mean parameters as explained earlier. It is observed that the scaled sparse coding over both the kinds of dictionaries does help in getting some improvements in recognition performance over that of the unadapted SI system.

3.4.3 Increasing the degrees of freedom

The presented SR-based approaches happen to substantially reduce the computational cost compared to jointly learning the scaling factors for all the K bases. At the same time, the restriction in the degrees of freedom in the weight estimation process does lead to some loss of information. It is quite evident from Figure 3.6(c) that the employed global scaling is bound to be suboptimal as there is a small spread about the mean. Consequently, the obtained recognition performance is noted to be inferior to that of the existing unconstrained ML estimation of the weights. To overcome this degradation, we have explored the possibility of increasing the degrees to freedom without much increase in the cost of weight estimation. To achieve the same, two kinds of dictionaries are employed in the sparse coding of the target. One of the dictionaries $\mathbf{\Phi}_1$ is the exemplar dictionary composed of speaker-specific supervectors. The eigenvectors derived using PCA done on the correlation matrix constructed from $\mathbf{\Phi}_1$ constitute the atoms of the other dictionary $\mathbf{\Phi}_2$. These dictionaries happen to provide two redundant representations of the same acoustic space. Next, the sparse coded targets $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$ are created using $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$, respectively. This is followed by the joint scaling of the two with SI pivoting to obtain the adapted model mean supervector as

$$\boldsymbol{\phi} = \bar{\boldsymbol{\phi}} + \widetilde{\mathbf{X}}\boldsymbol{\eta} \tag{3.22}$$

where $\widetilde{\mathbf{X}}$ is a matrix composed of the two sparse coded targets, i.e., $\widetilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1\ \tilde{\mathbf{x}}_2]$. The scaling factor vector $\boldsymbol{\eta} = [\eta_1\ \eta_2]^T$ is estimated using the iterative procedure given in Section 3.4.1.1 with (3.20) being modified as

$$\hat{\boldsymbol{\eta}} = \Bigg[\sum_{r=1}^{R}\Big(\sum_{l=1}^{L}\gamma^r(l)\Big)(\widetilde{\mathbf{X}}^r)^T(\bar{\mathbf{C}}^r)^{-1}\widetilde{\mathbf{X}}^r\Bigg]^{-1}\Bigg[\sum_{r=1}^{R}(\widetilde{\mathbf{X}}^r)^T(\bar{\mathbf{C}}^r)^{-1}\Big(\sum_{l=1}^{L}\gamma^r(l)[\mathbf{o}_l - \bar{\boldsymbol{\phi}}^r]\Big)\Bigg] \tag{3.23}$$

where $\widetilde{\mathbf{X}}^r$ is the part corresponding to the r^{th} Gaussian in $\widetilde{\mathbf{X}}$. We expect that an additional degree of freedom added due to separate sparse codings would help in checking the loss of

information to some extent.

One may argue that the need for two sparse codings would result in increased computational cost in the proposed approach. On the contrary, given the target, both the sparse codings can be performed in parallel. Thus, effectively the cost remains the same as that of single sparse coding. Increasing the number of dictionaries implies considering more number of bases in modelling which, in turn, may result in additional improvement in the adaptation performance. At the same time, with separate codings of the target, we are still able to contain the number of parameters required to be estimated. The complete details of the proposed approach is outlined in Algorithm 3.

3.4.4 Experimental evaluations and discussions

The setup used for the experimental evaluation remains the same as that discussed in Section 3.3. The performances for the proposed approach for the PFts test set with varying sparsity are given in Figure 3.7. The WERs are evaluated following the utterance-specific mode of adaptation. Different kinds of dictionaries, as discussed earlier, are explored and the results for the same are plotted. The performance is found to saturate when sparsity is around 10-14. The combination of the exemplar and learned dictionary happens to result in the least WER. This is so because this combination leads to capture more acoustic variations. Similar studies are done using the CAMts test set as well. Saturation in performance with a change in sparsity is noted in this case also. The best case results for these studies are tabulated in Table 3.5. For further studying the effectiveness of the proposed adaptation technique, we contrasted its recognition performance with other existing fast adaptation techniques like the EV and the RSW. The performances for the EV- and the RSW-based adaptation are also given in Table 3.5. It is to note that the recognition performance for the proposed approach employing two dictionaries is as good as the existing techniques.

In the case of the children's speech test set, the recognition performance of the SI system happens to be very poor. Consequently, we explored the feasibility of combining the proposed approach with the vocal tract length normalization. The VTLN-based frequency warping counters the ill effects of the variations in the vocal tract dimensions among the speakers. The differences in the vocal tract dimension is quite large between adult and child speakers. Consequently, the VTLN is found to be very effective in the case of the children's mismatched ASR [28]. To implement the same, warped MFCC features corresponding to frequency warping factors lying in the range of 0.88-1.12 and in steps of 0.02 are computed for the given test

Algorithm 3 SR-based Fast Adaptation Technique

Creation of sparse coded targets

Given: The exemplar the and learned dictionaries $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$, respectively and the SI model $\mathbf{\Lambda}$.

Step 1. Decode the test utterance using the SI model to generate the first-pass hypothesis.

Step 2. Perform mean only MAP adaptation of SI model using the test data and its first-pass hypothesis.

Step 3. Extract the mean supervector from the adapted model and substitute zeros for the unadapted dimensions to derive the *target*, $\mathbf{x}$.

Step 4. Using the OMP, derive the bases coefficients in SR $\boldsymbol{\upsilon}_1$ and $\boldsymbol{\upsilon}_2$ with respect to $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$ with the sparsity being set to K.

Step 5. Create sparse coded representations $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$ of the target using the sparse coefficients and the corresponding dictionaries

$$\tilde{\mathbf{x}}_1 = \mathbf{\Phi}_1 \boldsymbol{\upsilon}_1$$

$$\tilde{\mathbf{x}}_2 = \mathbf{\Phi}_2 \boldsymbol{\upsilon}_2$$

Joint estimation of scaling factors

Given: The matrix of sparse coded targets $\widetilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1\ \tilde{\mathbf{x}}_2]$, SI mean supervector, $\bar{\boldsymbol{\phi}}$, and SI covariance for r^{th} Gaussian, $\bar{\mathbf{C}}^r$, where $r = 1, \ldots, R$.

Step 6. Initialize scaling factor vector, $\boldsymbol{\eta}_1 = \mathbf{0}$

for $j = 1, j \leq \text{J}$, **do**

Step 7. Derive the mean parameter supervector for the adapted model

$$\boldsymbol{\phi}_j = \bar{\boldsymbol{\phi}} + \boldsymbol{\eta}_j^T \widetilde{\mathbf{X}}$$

if j = J **then**

Break

else

Step 8. Compute the posterior probability $\gamma^r(l)$ of the adaptation data for all r Gaussians in the adapted model under the constraint of the first-pass transcription

Step 9. Re-estimate the scaling factor vector

$$\hat{\boldsymbol{\eta}} = \left[\sum_{r=1}^{R}\Big(\sum_{l=1}^{L}\gamma^r(l)\Big)(\widetilde{\mathbf{X}}^r)^T(\bar{\mathbf{C}}^r)^{-1}\widetilde{\mathbf{X}}^r\right]^{-1}\left[\sum_{r=1}^{R}(\widetilde{\mathbf{X}}^r)^T(\bar{\mathbf{C}}^r)^{-1}\Big(\sum_{l=1}^{L}\gamma^r(l)[\mathbf{o}_l - \bar{\boldsymbol{\phi}}^r]\Big)\right] \tag{3.24}$$

end if

end for

Step 10. Estimate the final adapted mean supervector as

$$\hat{\boldsymbol{\phi}} = \bar{\boldsymbol{\phi}} + \widetilde{\mathbf{X}}\boldsymbol{\eta}_J$$

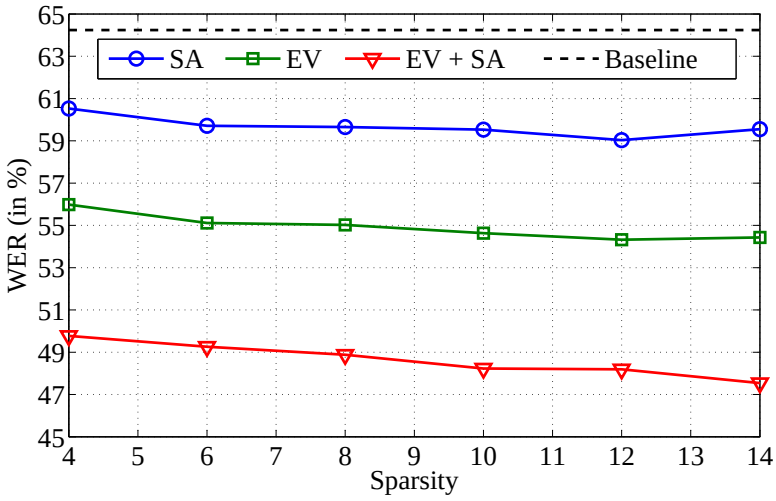

Figure 3.7: Recognition performances for the proposed scaled sparse coding-based adaptation technique on the PFts test set in the utterance-specific mode of unsupervised adaptation with varying sparsity. The sparse codings are performed on the exemplar (SA) and the learned (EV) dictionaries.

data. An ML-based grid search is then performed by aligning the different warped test features against the SI model under the constraints of the first-pass hypothesis to find the optimal warp factor. The optimally warped MFCC features are employed during the second-pass decoding on the SI models. Huge reductions in WERs due to the VTLN can be noted from Table 3.5. It is to note that, in the case of VTLN included baseline, the 95% confidence interval for the performance with respect to PFts is ± 1.30. Hence, the observed improvements in the recognition performances are statistically significant.

For amalgamating the power of the VTLN with the proposed approach, there exist two possible schemes:

- As the first-pass hypothesis is highly erroneous, one can first derive a better hypothesis using the adapted models. The grid search for the optimal warping factor is then done under the constraints of the second-pass hypothesis. A third-pass decoding using the warped MFCC features is performed next to obtain the final hypothesis. This incurs the latency of an extra decoding.

- Alternatively, the warping factors obtained using the first-pass hypothesis themselves could be used during the second-pass decoding with respect to the adapted model.

An absolute difference of 1.2% in WER for the above two schemes of applying the VTLN was noted with former being superior. In this work, for keeping the latency low we have followed the latter scheme. The combination of the VTLN and the proposed approach results in nearly

Table 3.5: The WERs for the proposed SR-based fast adaptation approach along with those of the existing methods in the utterance-specific mode. The WERs are shown for the adult test set (CAMts) as well as for the children test set (PFts). In the case of PFts, the WERs are given for the cases when the default as well as VTLN warped MFCCs are employed during decoding. The numbers in the parentheses denote the number of bases being interpolated to derive the adapted model parameters.

Explored Adaptation Techniques		WER (%)		
		CAMts	PFts	
			Default	VTLN
Unadapted SI		11.30	64.24	33.90
Existing	EV (16)	10.65	46.73	28.00
	RSW (18)	10.57	49.93	29.52
Proposed	1-Dictionary (1)	10.95	54.32	30.13
	2-Dictionary (2)	10.63	47.54	28.58

Table 3.6: The WERs for the proposed adaptation approach along with those of the existing fast adaptation methods in the incremental mode. The numbers in the parentheses denote the number of bases being interpolated to derive the adapted model parameters.

Explored Adaptation Techniques	WER (%)		
	CAMts	PFts	
		Default	VTLN
Unadapted SI	11.30	64.24	33.90
EV (16)	10.49	46.36	27.49
RSW (18)	10.50	50.26	29.88
Proposed (2)	10.51	47.19	28.05

additive improvement in the WER as shown in Table 3.5. We have also compared the proposed technique with existing approaches in the incremental mode of adaptation. The recognition performances for those experiments are tabulated in Table 3.6. Similar observations are noted in this case as well.

In the case of CAMts as well as PFts, significant improvements in the recognition performances are obtained by the use of the model-interpolation-based adaptation techniques. As already discussed, these techniques require the estimation of a much lesser number of parameters to derive the adapted models. This attribute makes them suitable for those applications where the adaptation data is generally scarce. The proposed adaptation approach performs similar to the mentioned fast adaptation techniques with only a few parameters being esti-

Table 3.7: The WERs of the proposed adaptation approach using regression-class-specific scaling factor in the utterance-specific mode for the PFts test. All the reported WERs include VTLN-based frequency warping being applied to the test features.

Explored Adaptation Technique	WER (in %)	
	Single reg-class	Two reg-classes
Unadapted SI	33.90	
EV	28.00	
Proposed (1-Dictionary)	30.13	28.62
Proposed (2-Dictionaries)	28.58	27.05

mated. Moreover, the consistency in improvement with increase in degrees of freedom from 1 to 2 can also be noted.

3.4.5 Regression class-specific scaling factors

Another way to increase the degrees of freedom is to partition the Gaussian components into classes as proposed in the CAT scheme. Instead of estimating a global scaling factor, class-specific scaling factors are estimated in this case. The Gaussian components can be partitioned into classes in several ways. In this work, we have clustered the Gaussians using the regression trees similar to that outlined in [49]. Increasing the degrees of freedom by using class-specific scaling factors helps in improving the recognition performance as evident by the WERs for the PFts test set given in Table 3.7. It is to note that, in the case of the exemplar dictionary, a relative improvement of 5% is obtained by increasing the number of classes to two. Further improvement can be obtained by using two dictionaries as evident from the last row in Table 3.7. In principle, we could employ higher number of regression classes. But that would lead to an enhanced overall computational cost as well as latency. Thus to limit our attention to low-latency fast adaptation tasks, we did not explore more regression classes for the scaling factor.

3.4.6 Contrast with recent works

A number of techniques, very similar to the EV approach, have been proposed in the past decade [56–59] as reviewed in Chapter 2. In all of these approaches, instead of estimating global interpolation weights, one interpolation weight parameter is estimated either for each segmental EV or for each of the feature dimensions [57,58]. Consequently, the number of parameters to be

estimated for adaptation is quite large. This, in turn, requires a greater amount of adaptation data for their robust estimation. These techniques are reported to outperform the EV approach when the adaptation data is significantly large.

A smoothing of ML weights using SR is proposed recently in [65]. In that work, initialization using ML weights and then smoothing using ℓ_1-regularization, significantly increases the complexity. This becomes prohibitive for the kind adaptation task dealt with in this work. In the SR-based adaptation approach reported in [67], the MP-based SR and ℓ_1-regularized SR are used. A comparatively better performances are obtained by using ℓ_1-regularized SR. For optimization in the case of ℓ_1-regularized SR, the projected gradient algorithm [69] is employed. The complexity of the projected gradient descent algorithm happens to be much higher than that of either OMP or LARS. In addition to that, it requires to be initialized using the coordinates of the bases selected using MP. A sequential ML criterion is used during sparse coding for optimization instead of the correlation-based representation error minimization used in the approach proposed in this work. Like the interpolation weight estimation, the complexity of the sequential ML optimization is square in the number of bases selected. Moreover, the ML optimization results in selecting the same atom more than once making the matrix of selected bases non-invertible. Use of OMP ensures that the selected bases are unique.

Recently, following the earlier proposed i-vector based adaptation approach, a fast adaptation technique for mobile speech task is reported [93]. In this work, a number of tools supporting parallel processing are employed to handle the large volume of data used in the training and the testing. But our unfamiliarity with those tools limits us to provide any meaningful contrast with the approach proposed in this work.

3.4.7 Discussion on the reduction of computational cost

The proposed scaled sparse coding technique greatly reduces the computations involved during model interpolation weight estimation since only a few scaling factors are to be estimated depending upon the number of dictionaries used. This reduction in number of weight parameters is achieved at the cost of performing multiple sparse codings of the target. Given that there are N atoms in the dictionary, the features are D dimensional and there are a total of R Gaussians, the complexity in the basis-search or the sparse coding using the OMP is $\mathcal{O}(DNR)$. The computational complexity is linear in each of the terms. On the other hand, if there are W weights to be estimated, complexity in 1-iteration of weight estimation is $\mathcal{O}(D^2WR + DW^2R)$. Reducing the number of weight parameters reduces the cost of estimation which depends as

Table 3.8: Comparison of the computational cost involved in the existing EV approach and the proposed sparse coding approaches employing a sparsity of 14 and a global scaling factor. Also given are the run-times in each case computed for a test file of about 8 seconds duration (847 frames).

Adaptation Technique	Sparse coding run-time† T_1 (secs)	Weight/scaling factor estimation	
		Number of parameters	Run-time† T_2 (secs)
EV	0	16	51
1-Dictionary	6	1	8
2-Dictionaries	12	2	12

† Computed for 64-bit MATLAB (R2010b) and HTK (ver. 3.4.1) running on Intel Xeon 6-core, 2.4 GHz CPU with 16 GB RAM.

the square of the number of weights. The proposed approach happens to significantly reduce the number of parameters. As already discussed, the multiple sparse codings can be done in parallel once the target supervector is derived from the given test data. Consequently, the complexity is effectively due to a single sparse coding.

To further substantiate our claim of reduction in computational cost, the run-times of these two steps are computed and enlisted in Table 3.8. It is to note that the times required to generate the first-pass hypothesis and the final decoding have not been considered as those remain the same for all the techniques. The value of T_1 is the sum of the time required to create the target supervector using MAP adaptation of SI model and to sparse code the target using OMP. Since EV does not require any basis-search or sparse coding, $T_1 = 0$ for that case. The value of T_2 is the time required to perform the ML estimation process to derive the interpolation weights or the scaling factors. The weight estimation process accounts for a major portion of the overall compute time as evident from Table 3.8. In the case of the EV, the value of T_2 is considerably higher than that for the proposed approach. Further, even the sum of T_1 and T_2 for the proposed technique is lower than that for the EV-based adaptation.

3.5 Summary

A novel fast adaptation approach employing scaled sparse coding over redundant dictionaries has been presented in this chapter. The proposed approach is found to result in a recognition performance similar to that obtained with the use of the existing techniques. The same is verified experimentally in this work using two different tests sets, viz. the adults' test set (CAMtr) and the children's test set (PFts). In the case of PFts, the proposed approach is also

combined with the VTLN to enhance the overall recognition performance. The combination of the VTLN with the proposed technique is found to result in additive improvements. Yet, the recognition performances observed for the mismatched testing case happen to be much poorer than those obtained in the matched case decoding. Large differences in the pitch values for the two groups of speakers happens to be one of the factors that severely degrades the performance in the mismatched testing case. Consequently, a study that attempts to quantify the reasons and the extent of pitch-induced distortions in the context of children's mismatched is presented in the next chapter.

4

Analysis of Pitch Induced Mismatch in Acoustic Features

Contents

As evident from the experimental results presented in Chapter 3, a highly degraded recognition performance is obtained when acoustic models trained on adults' speech are employed for decoding the children's speech. The acoustic properties such as formants, pitch, speaking rate, etc., vary significantly across adult and child speakers [21, 94]. These factors aggravate the acoustic mismatch between adults' and children's speech making the children's mismatched ASR a challenging problem. To compensate for these sources of mismatch, a number of techniques have been suggested in literature and were discussed in Chapter 1.

Among the various factors of mismatch, the differences in the vocal-tract dimensions between the adult and child speakers is a major one. Compared to the adults, the vocal organs of children are much smaller. These physiological differences account for a significant increase in the pitch as well as the formant values observed in the case of children's speech [21]. The acoustic mismatch resulting from the scaling of formant frequencies can be effectively addressed through the use of the VTLN. Consequently, the inclusion of the VTLN in the case of children's mismatched ASR results in large improvements as noted from the experimental results presented in Chapter 3. On the other hand, neither the VTLN nor the model-space adaptation approaches can effectively address the distortions induced due to the differences in the pitch values. The number of utterances in the CAMtr and the PFtr sets having a particular pitch value is plotted as a histogram shown in Figure 4.1. Similarly, the histograms for the pitch values for two test sets are shown in Figure 4.2. From these histograms, its clearly evident that the range of pitch values for the children's speech significantly greater than that for the adults' case.

In [40], it has been shown that the standard Mel-filterbank involved in MFCC feature computation is not able to provide sufficient smoothing especially for high-pitched child speakers. The insufficient smoothening of the pitch harmonics leads to a significant mismatch in the variances of the higher-order MFCCs for adults' and children's speech [41]. A number of studies have already been reported for addressing the pitch mismatch in the context of children's mismatched ASR [29, 30, 37, 40, 41, 95]. Significant improvements over the default modelling are reported either with synthetically transforming the pitch of the test signals to lower values [37, 41] or with suitable modification of a few lower-order filters in the Mel-filterbank during feature computation for the test signals [95]. These works happen to highlight the extent of the pitch-dependent mismatch that exists in such cases. In this chapter, we analyze in detail the cause and the extent of degradation induced due to the gross pitch mismatch. Based on the studies presented in this chapter, we will develop adaptation techniques specifically for

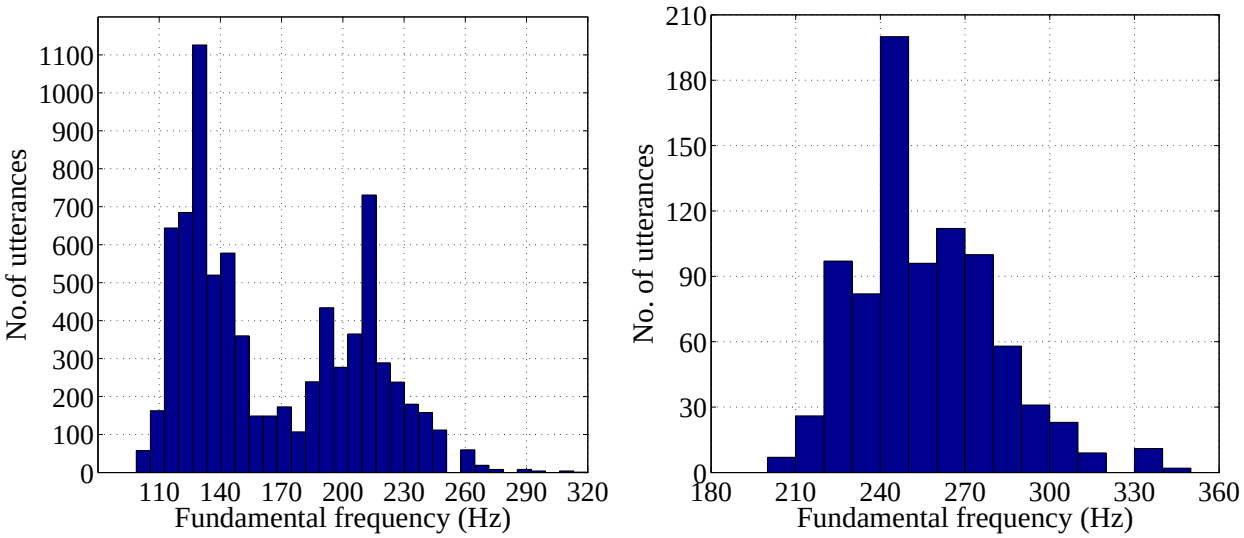

Figure 4.1: Histogram showing the number of utterances belonging to a particular pitch value in the CAMtr (left pane) and the PFtr (right pane) train sets, respectively.

children's mismatched ASR addressing the gross pitch mismatch in the following chapters.

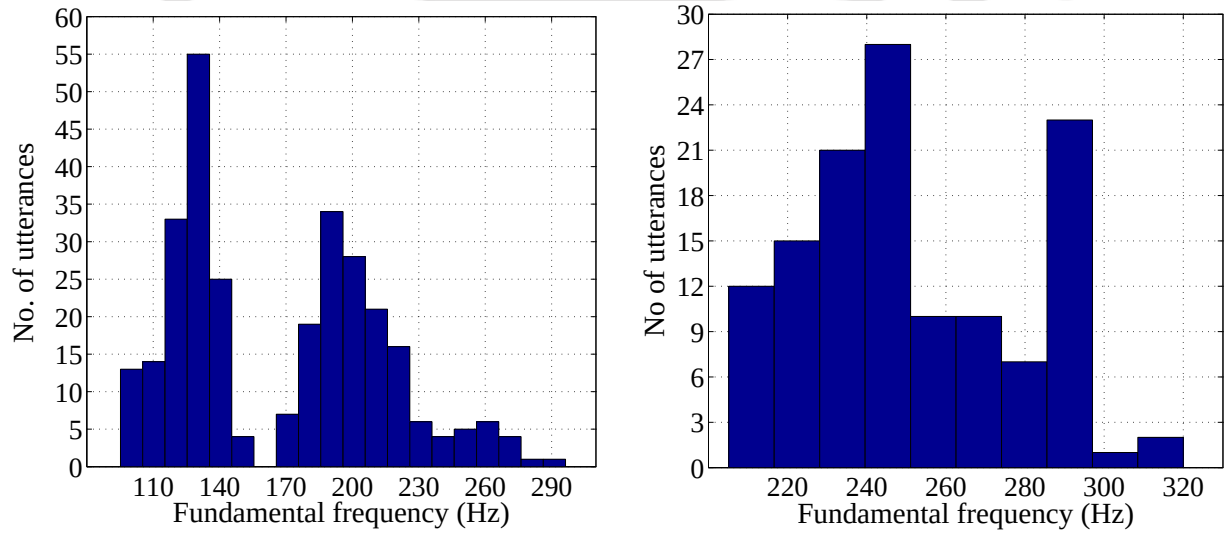

Figure 4.2: Histogram showing the number of utterances belonging to a particular pitch value in the CAMts (left pane) and the PFts (right pane) test sets, respectively.

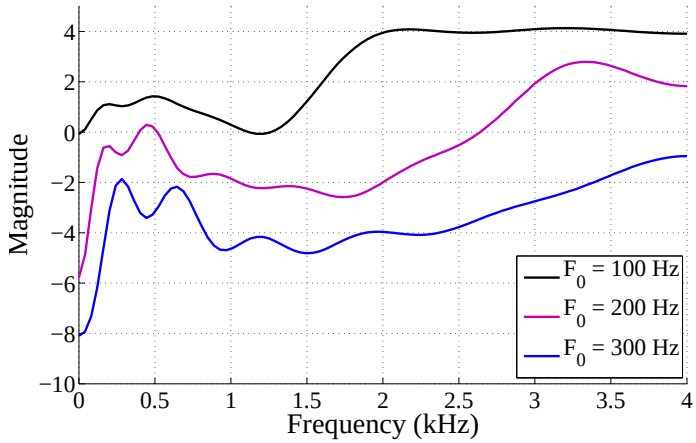

Figure 4.3: The spectral plots for the central frame of the vowel /IY/ is shown with variations in the F_0 value. The 13-dimensional base MFCC feature ($C_0 - C_{12}$) corresponding to the central frame are converted back to frequency domain using 100 point DFT. An intentional shift of 2 dB is added to make the curves distinguishable. The effect of pitch-dependent distortions is quite evident especially when $F_0 = 300$ Hz. This analysis is performed using the acoustic phonetic speech corpus TIMIT.

4.1 Need for Pitch Normalization

In ASR domain, the front-end parameterizations typically involve *static* signal processing, i.e., none of the parameters involved in the front-end signal processing algorithm are varied according to the characteristics of the signal being analyzed. For computing the commonly used MFCC features, first the spectral representation of the speech signal is obtained using short-time Fourier transform (STFT) analysis. The resulting spectrum is warped to non-uniform frequency scale and the same is effected by applying a triangular Mel-filterbank in the spectral domain. The filtered magnitude spectrum, after logarithmic compression, is then converted to the real cepstrum (RC) by applying discrete cosine transform (DCT). The low-time liftered version of the resulting cepstral coefficients yield the final MFCC features. Owing to low-time liftering, it is often assumed that such features would be largely free from the effect of excitation. On the contrary, earlier studies [28,40,41] have reported that the MFCC features do get affected for high-pitch (child) speakers in comparison to low-pitch (adult) speakers. It is argued that the periodicity of speech excitation is not preserved in the case of the non-uniform filtering unlike the uniform one. Thus, even with low-time liftering of the resulting Mel-cepstrum, some effect of the pitch would be present in the derived features. Further, during warping of the frequency scale, on account of narrow (≈ 100 Hz) bandwidth of lower-channel filters in the Mel-filterbank, the periodicity of the excitation is not well smoothed out while analyzing the signals having higher (> 200 Hz) pitch values. Consequently, some pitch-dependent distortions appear in the

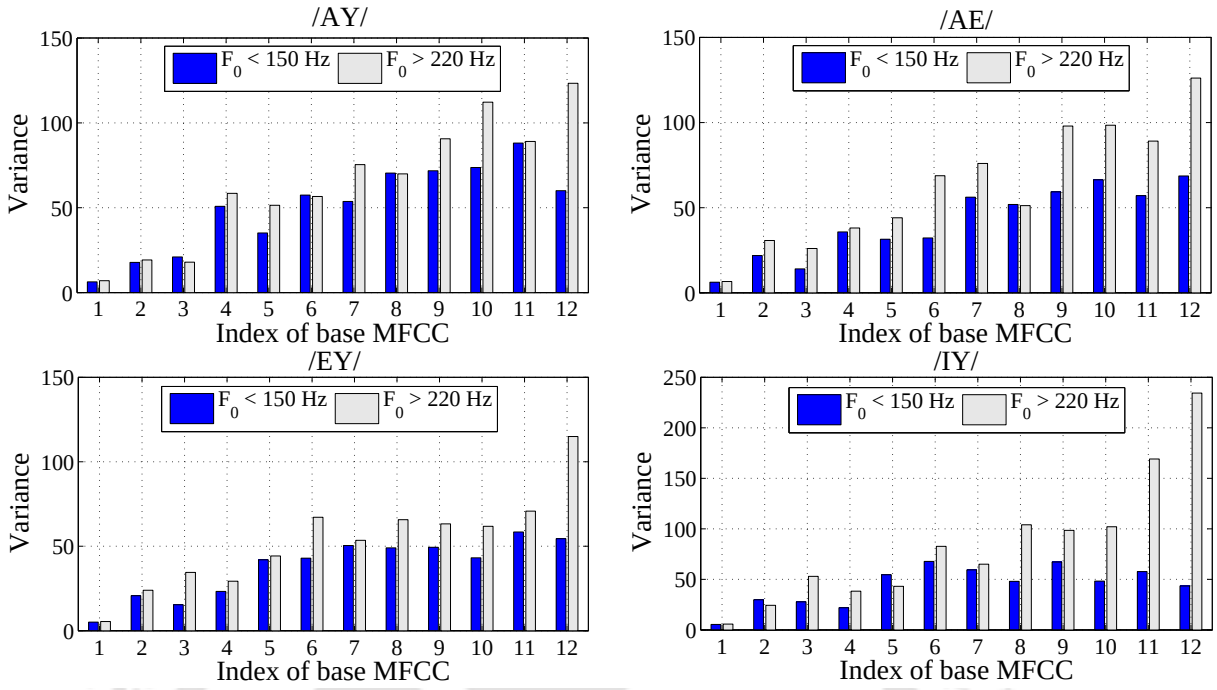

Figure 4.4: Variance plots for the base MFCC features C_1-C_{12} for four different vowels corresponding to two broad pitch (F_0) ranges. For this analysis, the feature vectors for nearly 2000 speech frames corresponding to the central portion of the vowel extracted from TIMIT are used. For the higher F_0 range, the mismatch in the variances of higher-order coefficients is evident.

lower frequency region of the spectral envelope for the children's speech. To demonstrate that, the smoothed spectral envelope corresponding to the base MFCC features for varying pitch values are plotted and are shown in Figure 4.3. As a result of the pitch- dependent distortions, the variance of the higher-order MFCCs increases for high-pitch speakers. To demonstrate that the dynamic range of higher-order MFCCs get enhanced in the case of the high-pitched speech, the variance of the base MFCC features of the vowels /AY/, /AE/, /EY/ and /IY/ for two broad pitch groups are plotted in Figure 4.4. In [96,97], salient other front-end cepstral features such as the linear prediction cepstral coefficient (LPCC), the perceptual LPCC (PLPCC) and the perceptual minimum variance distortionless response (PMVDR) were analyzed. Even those features were found to be sensitive to the variation in the average pitch values across adults and children.

4.2 Analytical Reasonings of Pitch Sensitivity of MFCC

In the following, we provide analytical explanation for the observed sensitivity of the MFCC features to varying pitch of speech signals with seeking answers to the following two queries:

(i) Why is the low-time liftering of the MFCC features no longer effective in removing the effect of the pitch?

(ii) How does the magnitude of higher-order coefficients in base MFCC features get affected by increasing the pitch?

From the *source-filter* model of speech production [98], the voiced speech signal is represented as convolution of the impulse response of the vocal-tract system $h(n)$ and the pulse train excitation $e(n)$. For voiced case, periodic excitation can be expressed as

$$e(n) = \sum_{q=-\infty}^{\infty} \delta(n - qP) \tag{4.1}$$

where P is the pitch period.
Then we have

$$\begin{aligned} s(n) &= h(n) * e(n) \\ &= \sum_{q=-\infty}^{\infty} h(n - qP). \end{aligned} \tag{4.2}$$

Consider a N-length frame of speech ending at time m and the same is defined as

$$\begin{aligned} f_s(n; m) &= s(n)w(m - n) \\ &= [h(n) * e(n)]w(m - n). \end{aligned} \tag{4.3}$$

Using the arguments of Oppenheim and Schafer [99], if the applied window $w(n)$ is long and tapers slowly compared to the time constants of $h(n)$, then we have

$$\begin{aligned} f_s(n; m) &\approx \sum_{q=-\infty}^{\infty} h(n - qP)w(m - qP) \\ &= h(n) * f_e(n; m) \end{aligned} \tag{4.4}$$

where $f_e(n; m) = e(n)w(m - n)$ denotes the frame of $e(n)$ determined by N-length window ending at time m.

In deriving the RC-based features for a speech frame, one seeks the short-term RC of the excitation and the long-term RC of the filter impulse response. From the approximation in (4.4), we have

$$c_s(n; m) = c_h(n) + c_e(n; m). \tag{4.5}$$

Already shown by Deller *et. al.* [100] (chapter 6, pp. 366-396) that for the causal voiced speech frames having rational bounded-input bounded-output (BIBO) stable system function, the RC of the filter impulse response (for being equivalent to the even part of the complex cesptrum) can be bounded as

$$c_h(n) \leq (Q^{in} + R^{in})\frac{\alpha^n}{n} + \frac{Q^{out}}{n}, \qquad n > 0 \tag{4.6}$$

where Q^{in} and Q^{out} represent the numbers of zeros inside and outside the unit circle, respectively, and R^{in} represents the number of poles inside the unit circle of the system function. The maximum modulus of the roots inside the unit circle is denoted by α. On the other hand, the RC of the excitation component can be expressed as

$$c_e(n;m) = \sum_{i=-\infty}^{\infty} \beta_i \delta(n - iP) \tag{4.7}$$

where β_is denote the impulse weights which decay as $1/i$ and depend on the window function used while P is the pitch period. Usually the envelope of $c_h(n)$ decays much quickly with respect to typical values of the pitch period P. Thus, an appropriate low-time liftering of $c_s(n;m)$ reduces the effect of the pitch.

In the MFCC feature computation, unlike the uniform RC derivation, the short-time spectrum is warped to a non-uniform Mel-scale prior to computing the RC. Now we attempt to qualify the effect of non-uniform frequency warping on the MFCC features. If the window used to derive the excitation frame covers Q periods of $e(n)$ and those periods corresponds to indices $q = q_0, q_0 + 1, \ldots, q_0 + Q - 1$, then

$$f_e(n;m) = \sum_{q=q_0}^{q_0+Q-1} w(m - qP)\delta(n - qP). \tag{4.8}$$

On taking non-uniform Fourier transform[1], we have

$$\begin{aligned} F_e(\omega_k;m) &= \sum_{n=m-N+1}^{m} f_e(n;m)e^{-j\omega_k n} \\ &= \sum_{n} \sum_{q=q_0}^{q_0+Q-1} w(m - qP)\delta(n - qP)e^{-j\omega_k n} \\ &= \sum_{q=q_0}^{q_0+Q-1} w(m - qP)e^{-j\omega_k qP} \end{aligned} \tag{4.9}$$

[1]In practice, the non-uniform discrete Fourier transform (DFT) is used with no preservation of the delay.

where ω_k denotes the frequency samples taken corresponding to the center frequencies of the triangular Mel-filterbank. Let us define the sequence

$$\tilde{w}(q) = \begin{cases} w(m - qP), & q = q_0, \ldots, q_0 + Q - 1 \\ 0, & \text{otherwise} \end{cases} . \tag{4.10}$$

Puting (4.10) into (4.9), we can show that

$$F_e(\omega_k; m) = \tilde{W}(\omega_k P) \neq \tilde{W}(\omega P) \tag{4.11}$$

where $\tilde{W}(\omega)$ denotes the uniform Fourier transform of the sequence $\tilde{w}(q)$. Therefore, due to the Mel-warping, the pitch harmonicity in the spectrum of the excitation frame is lost. On taking the DCT of $\log |F_e(\omega_k; m)|$ yields the Mel-cepstrum but it would not be pitch periodic. Thus, even if the same is low-time liftered, some effect of the pitch remains unlike the uniform RC case. This answers the first query.

Now for seeking the answer to the second query, we argue as follows. The spectra of the vocal filters being smoother, we can safely assume that even with the Mel-warping of the frequency axis the corresponding RC decays rapidly enough. Whereas, the Mel-cepstrum of the excitation frame being aperiodic, the bound on its envelope can be developed identically to the one derived for the RC of the impulse response part. Further, it is assumed that the functional form of that bound remains same as in (4.6). Under these assumptions, we argue that an enhanced envelope of the MFCC features would be noted in cases where the underlying rational systems representing the Mel-warped spectrum of the excitation frame happen to contain higher modulus roots within the unit circle. With increasing pitch values, more pitch dependent distortions would appear in the envelope of the Mel-spectrum of the excitation component. This will, in turn, lead to increased dynamic range of its RC. Therefore, the combined RC of the speech frame would exhibit enhanced dynamic ranges for the higher-order coefficients with increasing pitch of the signal. Here it is assumed that the magnitude of lower-order coefficients would be dominated by the RC of the filter impulse response.

For verifying the above arguments, we conducted experiments with synthetically generated periodic sequences and studied the nature of corresponding MFCC features and the pole-zero plots of the underlying minimum-phase systems[2]. Figure 4.5 shows the STFT magnitude spectra, the filtered Mel-spectra and the MFCC features derived for the synthetic excitation sequences corresponding to the pitch values of 100, 200, and 300 Hz. Further, the minimum-

[2]The MATLAB code for estimating the minimum-phase system from a non-minimum phase Mel-spectra is given in Appendix C.

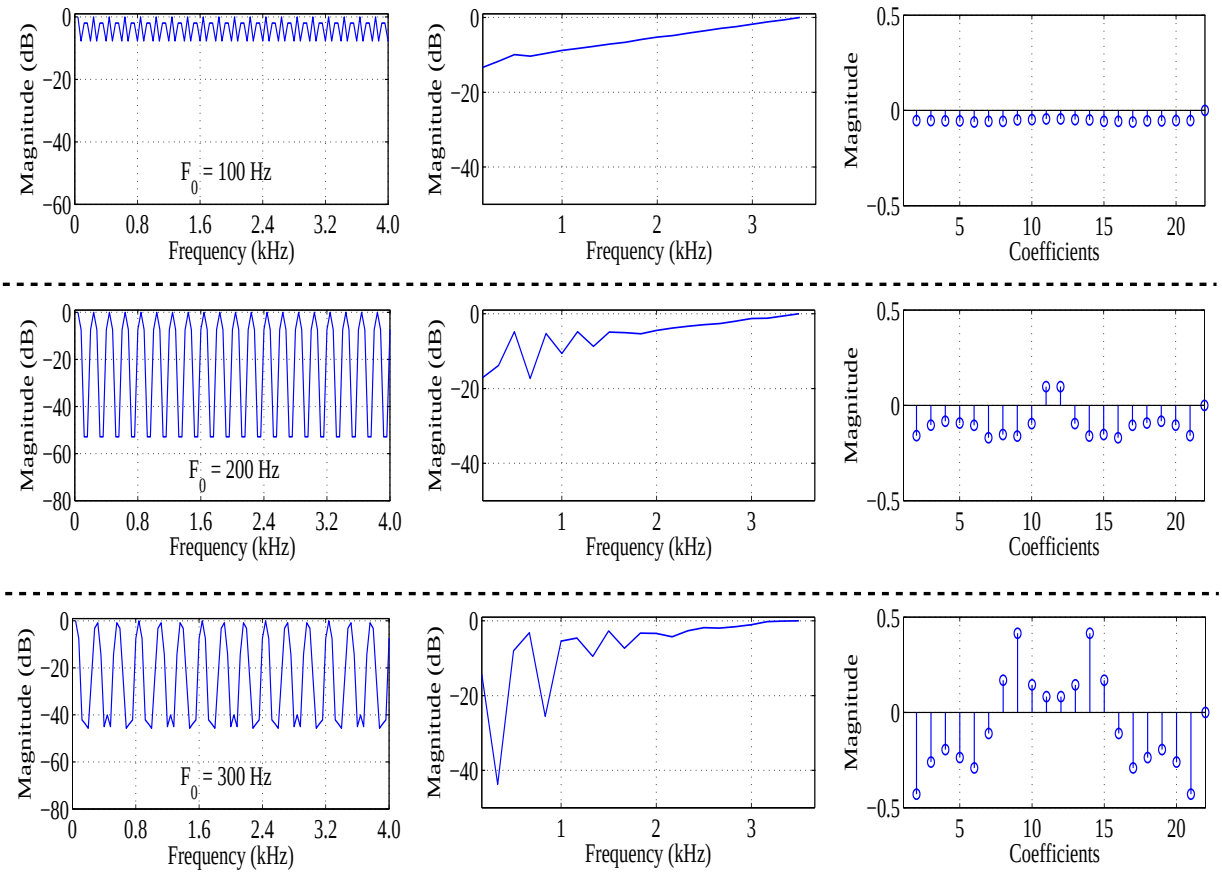

Figure 4.5: Plots demonstrating the increase in pitch dependent distortions in the Mel-spectral envelope as well as the increase in the magnitude of the corresponding cepstrum with the pitch of the excitation signals. The panels from left to right show the linear spectra, the filtered Mel-spectra and the real cesptra for the synthetically generated excitation frames, whereas the rows from top to bottom correspond to the pitch value of 100, 200, and 300 Hz, respectively.

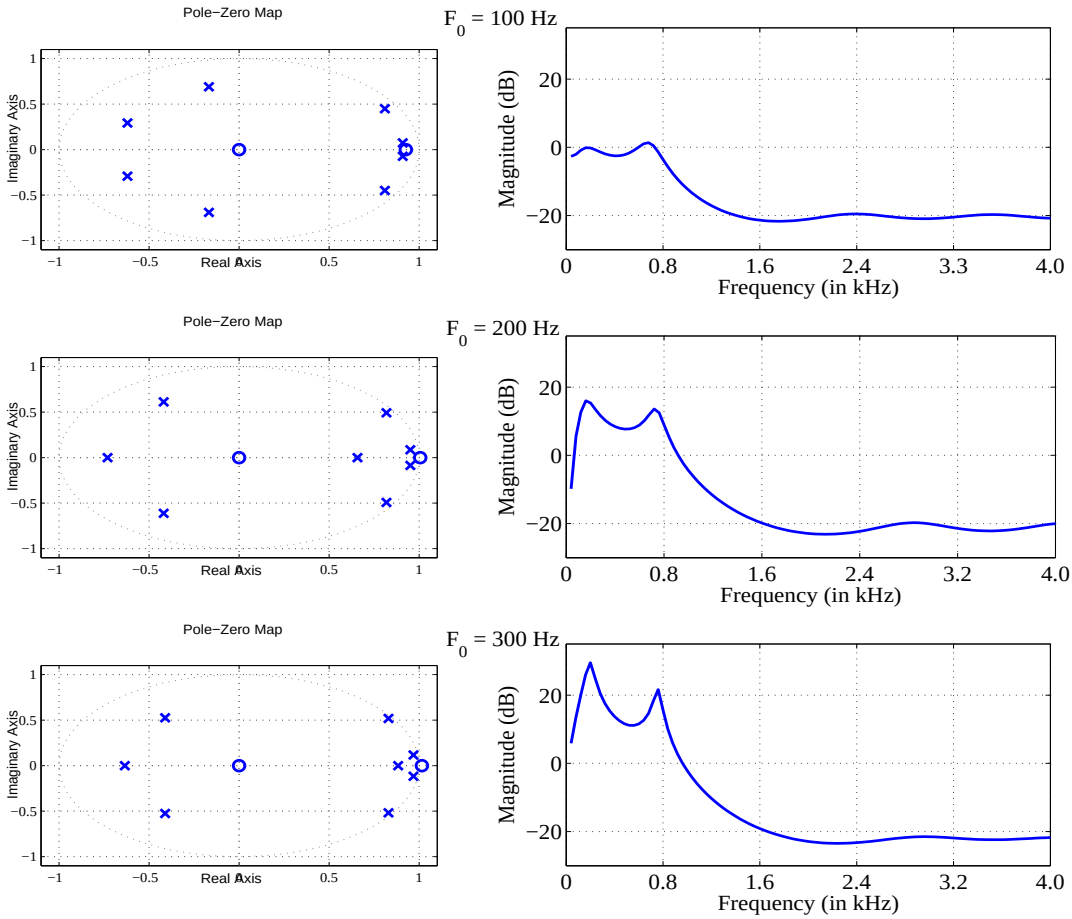

Figure 4.6: The pole-zero plots as well as the corresponding spectrum of the minimum-phase systems derived for varying pitch Mel-spectrum of the synthetic excitations shown in the middle panels of Figure 4.5. It can be noted that the moduli of the roots within unit circle are larger in the case of higher pitch values.

phase systems corresponding to the Mel-warped spectra are also derived and their pole-zero characteristics are plotted in Figure 4.6. From Figure 4.5, it can be noted that more pitch-dependent distortions appear in the Mel-spectral envelope with increasing values of the pitch even in the case of the synthetic excitation sequences having flat spectra. Also, the dynamic ranges of the corresponding cepstra enhance with increasing pitch values. On observing the pole-zero plots corresponding the minimum-phase systems, we can clearly note that the modili of the roots within the unit circle increase with increasing pitch distortion as argued.

4.3 Summary

When an ASR system is adults' speech employing the GMM-HMM-based acoustic modelling, the variances corresponding to the higher-order coefficients happen to be usually much smaller in comparison to the lower-order ones. On the other hand, the higher-order cesptral coefficients for the children's test speech exhibit much increased variance as discussed in the chapter. Consequently, on computing the likelihood of children's test data with respect to models trained on adults' speech, the Mahalanobis distance metric happens to enhance the distance score for the higher-order feature coefficients. This leads to a degradation in the likelihood which in turn degrades the recognition performance. To address this problem, a simple scheme that truncates some of the higher-order coefficients in the base MFCC features was explored in [40]. The enhanced spectral smoothening achieved by cepstral truncation helped in diluting the ill-effects of pitch-dependent distortions to a great extent. At the same time, it led to a significant loss of relevant spectral information. In the following chapters, we present alternate schemes to provide the needed spectral smoothening with a reduction in the loss of information.

Indian Institute of Technology Guwahati

5

Low-Rank Feature Projection-based Adaptation

Contents

To address the acoustic mismatch discussed in the previous chapter, a simple scheme that truncates some of the higher-order coefficients in the base MFCC features was explored in [40]. Although the BW scheme was noted to be quite effective, it leads to the loss of relevant spectral information when a large number of higher-order coefficients in the features are truncated out. To address the same, a structured low-rank projection-based soft-weighting (SW)[1] of features is proposed in this chapter. The primary objective of the SW approach is to map the adults' training speech and children's test MFCC features to a lower dimension subspace so that the mismatch in the variance is effectively reduced. For this purpose, a low-rank projection matrix is learned on the base MFCC features corresponding to adults' speech data. This can be done by employing either the principal component analysis (PCA) [55] or the hetroscedastic discriminant analysis (HLDA) [101]. Considering the nature of mismatch being similar in all three feature streams, the same projection matrix is applied to the delta and the delta-delta cepstral coefficients as well. Thus, the employed feature projection matrix has a *constrained block-diagonal* structure. The use of PCA/HLDA derived feature projections in ASR modelling is certainly not novel. We do come across HLDA-based projection being used for speaker adaptation [102]. But in that work, neither the employed projection was motivated for addressing pitch mismatch nor it had a block-diagonal structure. We have experimentally verified that such a constrained projection is more effective than the unconstrained one for the children's mismatched ASR.

In the second part of this chapter deals with the task of effectively combining the low-rank feature projections with some of the existing adaptation/normalization approaches. In this context, we propose a few novel structures which enable further reduction in the computational cost as well as the memory without much degradation in the adaptation performance. The salient contributions made in this chapter are summarized as follows:

- For addressing the acoustic mismatch between adults' and children's speech, a few structured low-rank feature projections are proposed.
- For further reduction of acoustic mismatch, the amalgamation of the soft-weighting with existing model-space-based adaptation approaches is explored.
- A novel adaptation approach that incurs a very low complexity and memory requirements is also proposed.

The remaining of the chapter is organized as follows: Section 5.1, we provide an overview of the binary weighting scheme. The low-rank feature projection techniques to address the

[1]soft-weighting (SW) and low-rank project will be used interchangeably in this thesis

acoustic mismatch are presented in Section 5.2. The proposed fast adaptation technique is described in Section 5.3. Finally, the concluding remarks are given in Section 5.4.

5.1 Motivation

As mentioned in the previous chapter, in adults' speech trained GMM-HMM-based acoustic models, the variances corresponding to the higher-order coefficients in the feature vector are usually much smaller relative to the lower-order ones. The mismatch in the variances of the features vectors corresponding to the adults' and children speech leads to a degradation in the likelihood which in turn degrades the recognition performance. To address this problem a simple BW scheme that truncates some of the higher-order coefficients in the base MFCC features was explored in [40]. The cesptral truncation can be formulated as a low-rank feature projection task. Let $\mathcal{H}_{K\times D}$ denote a projection matrix to be applied to a D-dimensional base MFCC feature vector such that $(D-K)$ higher-order coefficients are suppressed. The matrix in (5.1) depicts the case when only $K = 5$ lower-order coefficients are retained while the remaining higher order coefficients are forced to be zero.

$$\mathcal{H} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & \dots & 0 \end{bmatrix}_{5\times D} \tag{5.1}$$

Since, the velocity and the acceleration coefficients are also found to exhibit similar nature, the same transform is applied to those coefficients as well. The structure of the *constrained* projection matrix $\widetilde{\mathcal{H}}$ applied to the $3D$-dimensional test speech feature vector is given as

$$\widetilde{\mathcal{H}} = \begin{bmatrix} \mathcal{H} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathcal{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathcal{H} \end{bmatrix}_{3K\times 3D}. \tag{5.2}$$

For adaptation, the test data feature vectors as well as the mean vectors and the covariance matrices of the SI system are transformed using $\widetilde{\mathcal{H}}$. If $\mathbf{O} = [\mathbf{o}_1, \dots, \mathbf{o}_L]$ denotes the speech feature matrix consisting of L frames, the feature vector for the r^{th} frame after the projection is given by

$$\tilde{\mathbf{o}}_l = \widetilde{\mathcal{H}}\mathbf{o}_l. \tag{5.3}$$

Given the mean vector and the covariance matrix for the r^{th} Gaussian in the SI system ($\bar{\boldsymbol{\phi}}^r$ and $\bar{\mathbf{C}}^r$), respectively, the corresponding transformed mean vector and covariance matrix in the

adapted model are then given, respectively as

$$\tilde{\phi}^r = \widetilde{\mathcal{H}}\bar{\phi}^r \tag{5.4}$$

and

$$\widetilde{\mathbf{C}}^r = \widetilde{\mathcal{H}}\bar{\mathbf{C}}^r\widetilde{\mathcal{H}}^T. \tag{5.5}$$

Using the training data, the transformed means and covariances and $\widetilde{\mathcal{H}}$ as an input transform, a single iteration of maximum likelihood (ML) re-estimation is applied to optimize all the model parameters including the state transition matrices and the Gaussian mixture-weights.

5.2 Proposed Soft-Weighting Scheme

The drawback of the BW approach lies in the complete loss of information in the higher-order feature coefficients when those are dropped. Alternatively, one can learn the variation of the lower pitch (training) data and apply the same to the higher pitch (children's test) data. This scheme may be able to avoid the complete loss of information while suppressing the pitch-dependent mismatch. A low-rank projection capturing the principal dimensions of acoustic variations represented by the adults' speech training data is expected to suppress the higher-order coefficients. Thus, the dimensionality reduction techniques can be employed to learn such low-rank projections. In this work, we have explored the PCA and the HLDA for this purpose. Unlike the BW case, such low-rank projection matrices do not have a purely diagonal structure and therefore their application is referred to as *soft-weighting* of the cepstral features and the acoustic models.

5.2.1 Learning of structured low-rank projections

In speech signal processing, a number of dimensionality reduction techniques have been used for data compression and classification. One such approach is the principal component analysis. In the PCA, a new set of variables are defined that retain most of the information present in the original data variables. This is easily achieved by projecting the original data to a lower dimensional subspace capturing the maximum amount of variation. The bases spanning the desired lower dimensional subspace are called the *principal components* (PCs). The PCA-based projection ensures that the first PC has the largest variance. In other words, it accounts for the maximum variability in the data. Each succeeding component has the next highest variance. Furthermore, every succeeding component is orthogonal to the preceding PCs. The PCs are

derived by the eigen-decomposition of the *covariance* or the *correlation* matrix constructed out the given data.

Let $\mathbf{F}$ be matrix such that each column represents one of the $\mathcal{F}$-frames of the D-dimensional base MFCC feature vectors for the training set, $\{\mathbf{f}_i\}_{i=1}^{\mathcal{F}}$. The covariance matrix constructed out of $\mathbf{F}$ is given as

$$\mathbf{Q}_F = (\mathbf{F} - \bar{\mathbf{F}})(\mathbf{F} - \bar{\mathbf{F}})^T \tag{5.6}$$

where $\bar{\mathbf{F}}$ is the mean of $\mathbf{F}$. The correlation matrix, on the other hand, is given by

$$\mathbf{U}_F = \mathbf{F}\mathbf{F}^T. \tag{5.7}$$

If $\mathcal{H}$ denotes a $D \times D$ projection matrix, the transformed data matrix can be expressed as

$$\mathbf{V} = \mathcal{H}\mathbf{F}. \tag{5.8}$$

In the case of the PCA, the basic objective is to derive a $\mathcal{H}$ such that the covariance of the transformed data $\mathbf{Q}_V$ becomes diagonal. The eigen-decomposition of $\mathbf{Q}_F$ or $\mathbf{U}_F$ results in a set of D-dimensional eigenvectors $\{\mathbf{e}_i\}_{i=1}^{D}$ and their corresponding eigenvalues $\{\beta_i\}$. By selecting the eigenvectors corresponding to the K highest eigenvalues, dimensionality reduction with maximum retention of the information is achieved. These eigenvectors form the bases of the desired lower dimensional subspace. The $\mathcal{H}_{K\times D}$ is then arranged to form a block diagonal matrix as shown in (5.2) to derive $\widetilde{\mathcal{H}}$.

Another common dimensionality reduction technique is the HLDA. Unlike the PCA, class labels are required for learning the HLDA-based low-rank projection. Given the labeled training data, an ML approach minimizing the ratio of the within-class and the between-class scatter is used to derive the HLDA-based projection matrix. In order to derive $\widetilde{\mathcal{H}}$ in the structure given in (5.2), a GMM-HMM-based system is developed using the D-dimensional base MFFC features only. A projection matrix $\mathcal{H}_{K\times D}$ is then learned using the developed GMM-HMM system and the labeled adults' training data. Since the projection matrix in the PCA case is not learned in ML sense, it is not expected to be as effective as that in the HLDA case.

In order to depict the degree of suppression graphically, a 13×13 identity matrix $\mathbf{I}$ is used to study the effect of projection on a particular dimension of the feature vector. Each column of $\mathbf{I}$ represents one of the 13 dimensions in the feature vector being high while the remaining 12 dimensions being deactivated. The projection matrix $\mathcal{H}_{K\times D}$ is then multiplied with $\mathbf{I}$. For each of the columns in the transformed matrix, the ℓ_2-norm is computed to determine the resulting energy. The dimension-wise resulting energy for the three kinds of projections (BW, SW-PCA

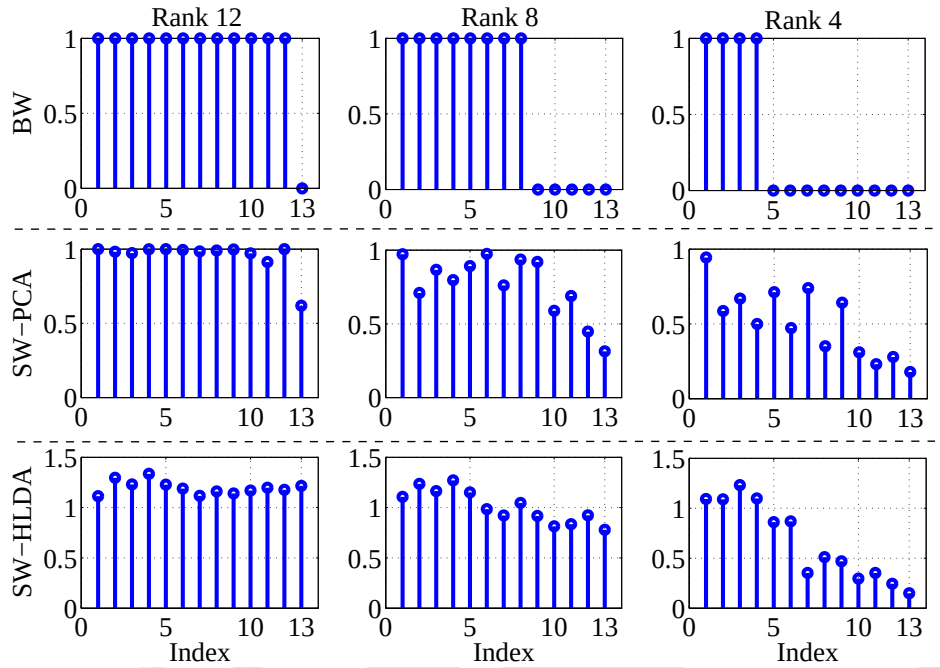

Figure 5.1: Feature dimension-wise energy distributions obtained by multiplying $\mathcal{H}_{K\times D}$ with a 13×13 identity matrix $\mathbf{I}$. The resulting energies are shown for the three different kinds of projection matrices with varying rank ($K = 12$, $K = 8$ and $K = 4$). In these plots, the x-axis denotes the feature coefficient index while the y-axis represents the magnitude of the resulting energy. These plots highlight the degree of suppression of higher-order indices in the feature vector in each of the cases.

and SW-HLDA) are shown in Figure 5.1. These distributions are plotted for the cases when $K = 12$, 8 and 4. As both the variants of PCA are observed to have a similar effect, the energy distribution for the correlation-matrix-based PCA case is plotted here. With reducing rank of the projection matrices, an increased suppression of higher-order coefficients is noticeable for all three cases. But unlike the BW case, some information in the higher-order coefficients would be retained in the SW cases which will lead to an enhanced recognition performance.

5.2.2 Performance evaluation

For evaluating the proposed low-rank projections, the rank of $\widetilde{\mathcal{H}}$ is varied from 36 to 12 in steps of 3. For incorporating the VTLN, the warped feature vector $\mathbf{o}_l^{(\alpha_i)}$, α_i being the optimal warp factor, is transformed by $\widetilde{\mathcal{H}}$. The transformed feature vector $\tilde{\mathbf{o}}_l^{(\alpha_i)}$ is then decoded on the correspondingly transformed models. The WER profiles for the SW-PCA-based feature projection technique with variations in the rank of the applied projection matrices are shown in Figure 5.2(a). For contrast, the WERs for the case when unstructured projection matrices are employed are also shown in Figure 5.2(a). For deriving the unstructured projection matrices,

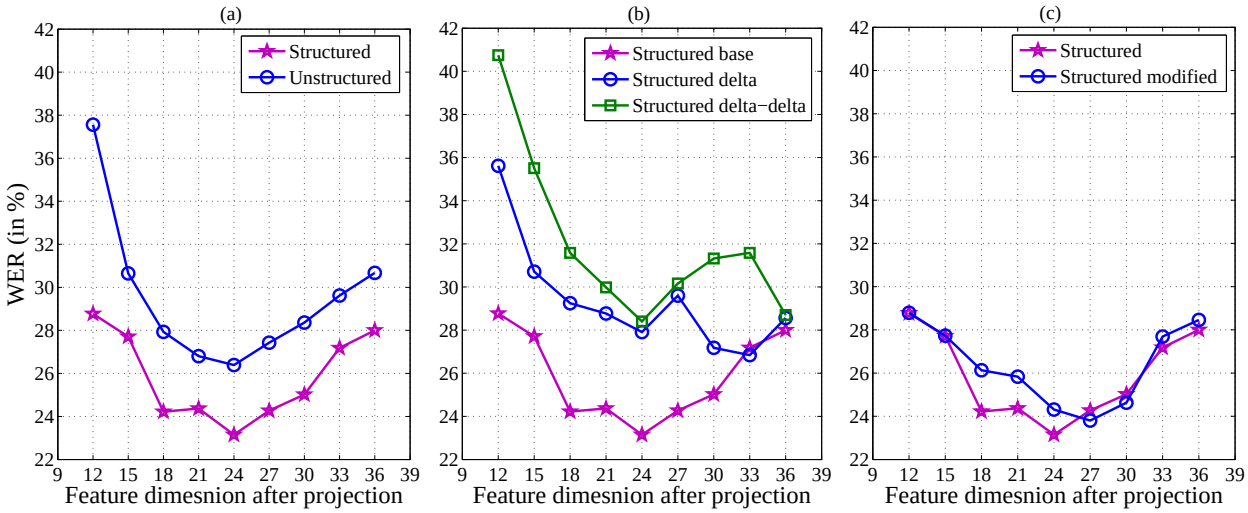

Figure 5.2: The WER profiles for children's mismatched ASR for varying ranks of SW-PCA with (a) structured and unstructured feature projections, (b) root projection matrix $\mathcal{H}$ derived using base, delta and delta-delta features, respectively, and (c) modified structured feature projection.

the 39-dimensional MFCC feature vectors are used during the PCA. The efficacy of employing the structure given in (5.2) is evident from Figure 5.2(a). A possible explanation for the use of proposed structure lies in the fact that both the delta and the delta-delta coefficients in the feature vector have much lower variances in comparison to those of the base coefficients. Consequently, in the case of the unstructured projection, mostly the delta and the delta-delta coefficients would get suppressed with a reduction in the rank instead of an effective suppression of only the higher-order coefficients in all three streams. Furthermore, the derivation of the root projection $\mathcal{H}$ using the delta/delta-delta coefficients is also explored and the WER profiles for the same are shown in Figure 5.2(b). An absolute degradation in WERs by 3.5% and 5.5% are noted.

Another way to derive the structured low rank projection is as follows: derive $\mathcal{H}_1$ using the base, $\mathcal{H}_2$ using the delta and $\mathcal{H}_3$ using the delta-delta coefficients, respectively. The structure of the *modified* projection matrix $\widetilde{\mathcal{H}}$ applied to the $3D$-dimensional test speech feature vector is then given as

$$\widetilde{\mathcal{H}} = \begin{bmatrix} \mathcal{H}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathcal{H}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathcal{H}_3 \end{bmatrix}_{3K \times 3D} . \tag{5.9}$$

We explored the effect of employing the modified structure as well. The WER profile for this study along with that for the original structure given in (5.9) with variations in the rank are shown in Figure 5.2(c). Interestingly, both the structures result in similar changes in the

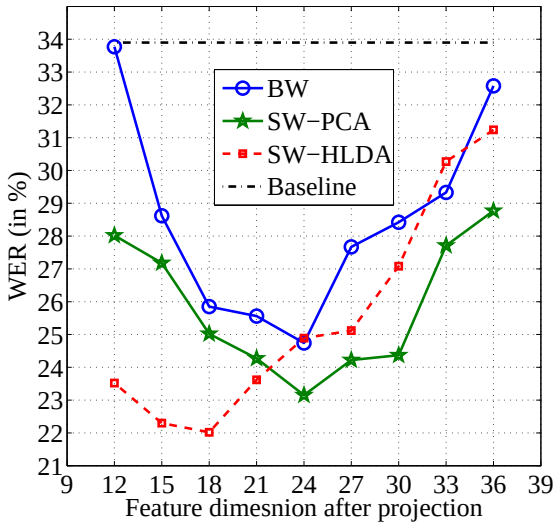

Figure 5.3: The WER profiles with varying ranks of the BW and the SW-based feature projections for children's mismatched ASR.

WER with the variation in the rank of the projection matrix. For the ease in comparison, the WER profiles for three kinds of projections (BW, SW-PCA and SW-HLDA) are shown in Figure 5.3. It can be noted that the proposed SW-based approach significantly outperforms the BW approach. It is to note that the the 95% confidence interval for the performance with respect to PFts is ±1.30. Therefore, all the observed improvements in the recognition performances are statistically significant.

To further build the confidence in our proposed approach, another ASR system was developed using the PLP-based acoustic features instead of the MFCC features. The procedure of deriving the PLP-based cepstral features is described in Appendix A.2. 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 that reported in the literature. 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 so derived 39-dimensional features are employed for the training of the ASR system. The developed ASR system had the same architecture as discussed in Chapter 3. The recognition performance of new SI system for the PFts test set turns out to be very similar to that obtained using the MFCC features. Furthermore, the use of VTLN and low-rank feature projection is observed to be effective with the PLPCC features as well. The WER profile for this study with variations in the rank of the SW-HLDA-based projection matrix are shown in Figure 5.4. For the contrast purpose, the WER profile for the MFCC

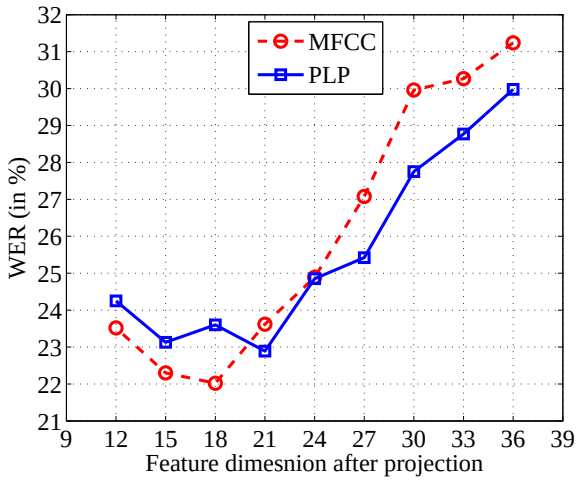

Figure 5.4: The WER profiles with varying ranks of the BW and the SW-based feature projections for children's mismatched ASR.

features are also shown in that figure.

5.3 Revisiting Fast Adaptation of Acoustic Models

In the previous chapters, the model-interpolation-based fast adaptation techniques were discussed in detail. Moreover, their efficacy was explored under the matched and the mismatched case decoding. In the following, we first re-present an overview of the fast adaptation techniques for quick reference. This is followed by a discussion on a novel fast adaptation technique that significantly reduces the latency in comparison to the existing approaches.

In the model-interpolation-based fast adaptation techniques, the basic assumption is that the adapted Gaussian mean vectors lie in a low dimensional subspace. The adapted GMM-HMM means are hence derived by a linear combination of a set of predefined acoustic models (or *bases*). For example, the adapted mean vector for the r^{th} Gaussian $\boldsymbol{\phi}^r$ is derived as

$$\boldsymbol{\phi}^r = \boldsymbol{\phi}_0^r + \eta_1 \boldsymbol{\phi}_1^r + \ldots + \eta_j \boldsymbol{\phi}_j^r + \ldots + \eta_M \boldsymbol{\phi}_M^r \tag{5.10}$$

where $\boldsymbol{\phi}_0^r$ is an optional bias vector and $\{\boldsymbol{\phi}_j^r\}_{j=1}^M$ is the set of predefined bases. Since the other GMM-HMM parameters remain unchanged, those are borrowed from the SI system to define the complete model. Only a few interpolation weights or the *direction coordinates* $\{\eta_j\}$ are required to be estimated. These interpolation weights are global in nature, i.e., the same weight is applied to all the Gaussians in a basis. As a consequence of these constraints, even a small amount of adaptation data happens to be sufficient for their reliable estimation. In the

case of ASR tasks involving human machine interactions, such a reduction in the number of parameters is highly desirable.

The complexity of the interpolation weight estimation process depends on a number factors. If R is the number of Gaussians in the acoustic model, M is the number of bases and D is the dimensionality of MFCC features, then the complexity of weight estimation is $\mathcal{O}(R[M^2D + MD^2])$. When the number of bases considered is not large ($M \leq D$), the complexity can be approximated as $\mathcal{O}(R\ M\ D^2)$. The cost of MLED can be reduced by reducing the value of M. Exploiting this fact, we had presented an SR-based fast adaptation technique in Chapter 3. The computational cost of MLED can be further reduced by reducing D. Consequently, we explored the feasibility of combining the SW-based projection techniques with model-interpolation-based adaptation approaches. Such reduction is highly desirable in the case of ASR tasks where latency is major factor as well as the available adaptation data is low. Moreover, this is also expected to result in additive improvements in the system performance.

5.3.1 Inclusion of soft-weighting in fast adaptation

In general, the adults have larger vocal organs than the children. Consequently, the adults' speech has lower pitch and formant values. As a child follows a normal growth profile, the diameter of glottis and the length of vocal tract remain in proportion. With this assumption, the pitch value and the vocal tract length are expected to have a strong negative correlation. Thus, both the pitch and the vocal tract length differences happen to contribute additively to the acoustic mismatch observed between adults' and children's speech. Based on these arguments, a linear dependence between the degree of acoustic mismatch and the VTLN warp factor for the test data estimated with respect to the trained acoustic model was explored in [28]. Through an empirical study done using the PFtr set in the PF-STAR database, it was shown that the rank of the BW transform to be applied for a particular speaker could be selected based on the estimated VTLN warp factor for that speaker. We repeated that study for the SW-based projection and observed a similar linear dependence.

In order to derive the lookup table, a development set is created from the PFtr, i.e., the PF-STAR training data. The development set consists of 350 utterances from 66 speakers constituting a total of 16560 words. Care has been taken to include the utterances from speakers belong to all age groups (4-13 years). The WER for the development set on the SI system without and with the inclusion of the VTLN happen to be 54.96% and 39.50%, respectively. The split of the development set on the basis of the optimal VTLN warp factor

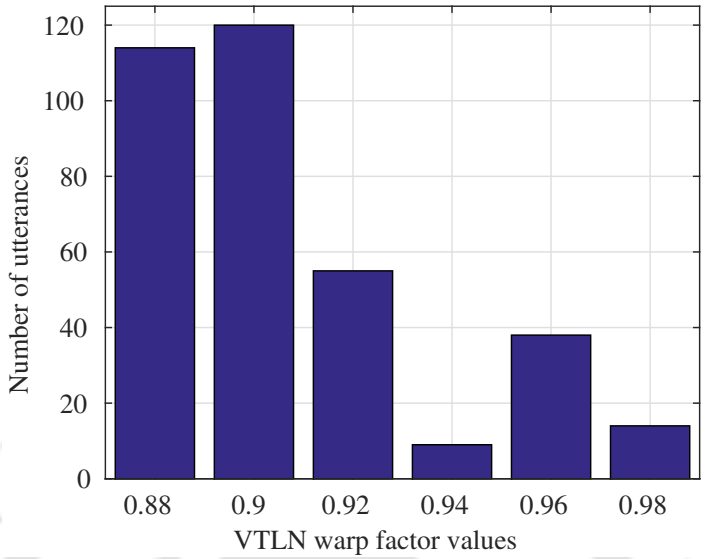

Figure 5.5: Distribution of utterance-specific VTLN warp factors estimated with respect to the SI system for a development set consisting of 350 utterances extracted from the PFtr.

values estimated using the SI systems is shown in Figure 5.5. The WERs for each of the split in the development data with respect to differently SW-PCA transformed models is shown in Table 5.1. Using a linear fit to best case WERs obtained in this study, the lookup table is determined. Thus, given the VTLN warp factor for the children's test data, the rank of the projection to be applied is selected as per the look-up table shown in Table 5.2.

During the system development phase, the SI system $\mathbf{\Lambda}$ is transformed by applying different low-rank projection matrices $\widetilde{\mathcal{H}}^{(\alpha_i)}$ corresponding to warp factor (α_i) values lying in the range 0.88 to 0.98 in steps of 0.02. The chosen range corresponds to the warp factor values that are usually found to be optimal in the case of children's mismatched ASR. Let the transformed SI system be denoted by $\mathbf{\Lambda}^{(\alpha_i)}$. The speaker-specific training data in CAMtr is also transformed using various $\widetilde{\mathcal{H}}^{(\alpha_i)}$'s. The transformed speaker-specific training data is then used to adapt the Gaussian mean vectors of the corresponding $\mathbf{\Lambda}^{(\alpha_i)}$. A separate set of speaker-specific mean supervectors are then extracted from each mean adapted $\mathbf{\Lambda}^{(\alpha_i)}$. Next, a separate set of eigenvectors are derived from those speaker-specific mean supervectors corresponding to each α_i following the procedure explained earlier. In the case of EV-based adaptation, for each each α_i, the top M eigenvectors form the bases. Similarly, in the case of SR-based adaptation, for each α_i, two dictionaries are created using the speaker-specific mean supervectors and the eigenvectors, respectively.

During the mismatched testing, the value of α_i is computed for the given test data under the

Table 5.1: WERs for the optimal warp factor based splits of the development set with respect to various SW-transformed models.

Rank of the SW matrix	WER (in %)					
	0.88	0.90	0.92	0.94	0.96	0.98
12×13	37.41	39.89	20.73	13.94	13.17	**15.63**
11×13	36.52	38.83	20.05	**13.05**	**11.94**	16.08
10×13	33.97	37.84	19.53	15.04	13.40	17.15
9×13	34.58	37.46	**19.23**	14.38	13.17	16.84
8×13	33.23	**36.30**	20.05	16.15	13.73	16.39
7×13	**32.93**	40.22	23.28	15.49	15.41	18.06
6×13	37.45	39.72	21.36	17.04	16.87	18.51
5×13	38.02	44.71	25.49	17.48	21.52	22.61
4×13	42.73	46.64	30.13	22.35	27.19	24.89

Table 5.2: A look-up table for selecting the order of the feature projection matrix $\widetilde{\mathcal{H}}^{(\alpha_i)}$ based on the VTLN warp factors $\{\alpha_i\}_{i=1}^{6}$.

Warp factor value	Order of the projection matrix	Warp factor value	Order of the projection matrix
$\alpha_1 = 0.88$	21×39	$\alpha_4 = 0.94$	30×39
$\alpha_2 = 0.90$	24×39	$\alpha_5 = 0.96$	33×39
$\alpha_3 = 0.92$	27×39	$\alpha_6 = 0.98$	36×39

constraints of the first-pass hypothesis. Depending upon the chosen adaptation technique, the corresponding set of eigenvoices or the dictionaries are selected using the look-up table (shown in Table 5.2). In the case of EV, the adapted model mean vector for the r^{th} Gaussian is then derived as

$$\hat{\boldsymbol{\phi}}^r = \mathbf{e}_0^{r,(\alpha_i)} + \sum_{j=1}^{M} \eta_j \, \mathbf{e}_j^{r,(\alpha_i)} \tag{5.11}$$

where the interpolation weights η_j's are estimated using MLED. On the other hand, in the case of SR-based adaptation, a target supervector $\mathbf{x}^{(\alpha_i)}$ is created using the given test utterance transformed using $\widetilde{\mathcal{H}}^{(\alpha_i)}$. The target is sparse coded over the two selected dictionaries to derive $\tilde{\mathbf{x}}_1^{(\alpha_i)}$ and $\tilde{\mathbf{x}}_2^{(\alpha_i)}$, respectively. The adapted model mean vector for the r^{th} Gaussian in this case

Table 5.3: The WERs for the EV- and the SR-based fast adaptation approaches implemented in the utterance-specific mode. The number in parenthesis indicates the bases employed in that case. The performances are also shown for the cases when the SW-based projections are combined with the EV- and the SR-based adaptation.

Explored adaptation techniques	WER (in %)
No adaptation	33.90
EV (16)	28.00
SR (2)	28.58
SW-PCA	23.15
SW-PCA + EV (16)	19.83
SW-PCA + SR (2)	20.09
SW-HLDA	22.02
SW-HLDA + EV (16)	18.24
SW-HLDA + SR (2)	18.69

is then derived as

$$\hat{\boldsymbol{\phi}}^r = \mathbf{a}_0^{r,(\alpha_i)} + \eta_1 \, \tilde{\mathbf{x}}_1^{r,(\alpha_i)} + \eta_2 \, \tilde{\mathbf{x}}_2^{r,(\alpha_i)} \tag{5.12}$$

The study of including SW projections in model interpolation is performed on PFts test set only. The WERs for the same are given in Table 5.3. A relative improvement of 14% and 21% over the EV adaptation are noted for the PCA- and the HLDA-based projections, respectively. Similar reductions in WER are obtained with the SR-based adaptation as well. Since the the 95% confidence interval for the performance with respect to the SW baseline is ± 1.16, the observed improvements in the recognition performances are statistically significant. It is to note that the warp factors for a large number of the test utterances in PFts test set lie in the range of 0.88-0.92. Consequently, the employed MFCC features and the model parameters happen to be 21-27 dimensional. Since the complexity of MLED is quadratic in feature dimension, a substantial reduction in the cost is achieved.

The adaptation scheme outlined in this section requires extra memory for the storage of M eigenvoices for each of the warp factor values in contrast to an unadapted SI system. In the case of the SR-based approach, the memory requirement is even more since all the mean supervectors as well as the eigenvectors are required to be stored. Such an increase in memory requirement may not be amenable for interactive ASR tasks. In the next section, we describe

a novel approach for addressing both these issues.

5.3.2 Proposed fast adaptation approach

In the proposed technique, a set of Gaussian mean supervectors are first derived *a priori* using some data from children domain[2]. These mean supervectors can be derived using either the EV- or the MLLR-based adaptation of the SI system. In this study, the SR-based approach is not used because of the increased memory requirements. Given the test data, a mean supervector is selected and then optimally scaled to get the adapted mean vectors. Since a single scaling factor is required, the introduced latency is minimal. In the following, the two steps involved in proposed adaptation approach are discussed.

5.3.2.1 Off-line estimation of model mean parameters

The children's speech training data (PFtr) is treated as the domain data and is split into different acoustic groups corresponding to their VTLN warp factors estimated with respect to the SI system $\mathbf{\Lambda}$. In the case of EV, for each α_i, an interpolation weight vector $\boldsymbol{\eta}^{(\alpha_i)}$ is derived via MLED using the corresponding domain-specific data and its true transcription. The mean vector for the r^{th} Gaussian is then estimated as

$$\widehat{\mathbf{m}}^{r,(\alpha_i)} = \mathbf{e}_0^{r,(\alpha_i)} + \sum_{j=1}^{M} \lambda_j^{(\alpha_i)} \mathbf{e}_j^{r,(\alpha_i)}. \tag{5.13}$$

Similarly, in the case of MLLR, given the domain-specific data corresponding to an α_i and its true transcription, a global affine transform of the SI means is computed. For the r^{th} Gaussian in the model $\mathbf{\Lambda}^{(\alpha_i)}$, the transformed mean vector is estimated as

$$\widehat{\mathbf{m}}^{r,(\alpha_i)} = \mathbf{P}^{r,(\alpha_i)} \bar{\boldsymbol{\phi}}^{r,(\alpha_i)} + \mathbf{b}^{r,(\alpha_i)} \tag{5.14}$$

where $\mathbf{P}^{r,(\alpha_i)}$ is the MLLR transform, $\bar{\boldsymbol{\phi}}^{r,(\alpha_i)}$ is the SW transformed SI mean vector corresponding to the value of α_i and $\mathbf{b}^{r,(\alpha_i)}$ is a fixed bias. The discussed procedure for creating the mean vectors is outlined as a block diagram in Figure 5.6. In both these cases, only the transformed mean vectors $\mathbf{m}^{r,(\alpha_i)}$ are saved. One can stack the transformed mean vectors $\mathbf{m}^{r,(\alpha_i)}$ in the form of a supervector $\mathbf{m}^{(\alpha_i)}$ as well before saving.

During testing, the pre-estimated mean vectors are chosen on the basis of the value of α_i estimated for the test data. The remaining model parameters used in decoding are borrowed

[2]This scheme is feasible only when a small amount of developmental data from the children domain is available. In the absence of the domain data, the approach described in Section 5.3.1 may be used.

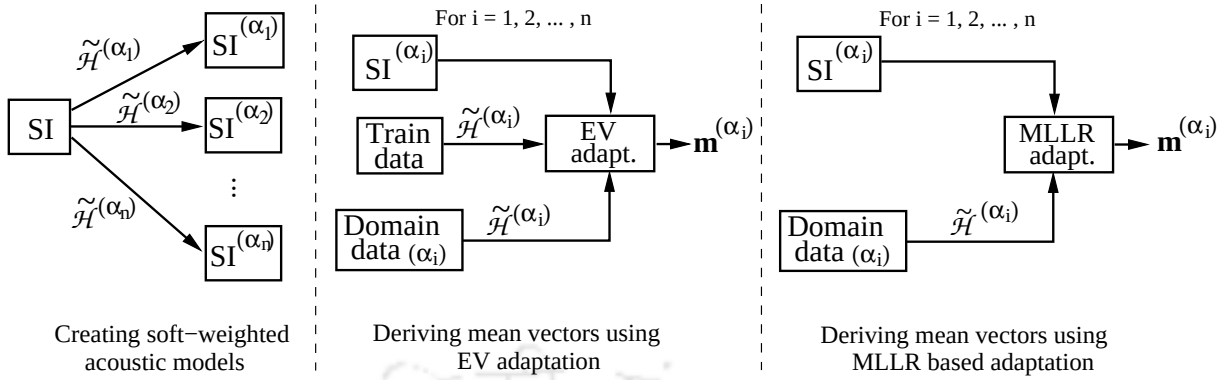

Figure 5.6: The block diagram details the steps involved in creating the transformed mean vectors from the SI system using a combination of SW and model space-based adaptation. The process is outlined for the two different adaptation approaches, viz. Eigenvoices (EV) and maximum likelihood linear regression (MLLR). $\widetilde{\mathcal{H}}^{(\alpha_i)}$ denotes a structured low-rank projection matrix whose order is tied to the VTLN warp factor α_i as shown in Table 5.2.

from the corresponding SW transformed SI system $\mathbf{\Lambda}^{(\alpha_i)}$. By employing the VTLN warp factor-based indexing of models, a single parameter α_i is required to be estimated for the given test data. With this scheme, the entire process of deriving the model parameters becomes off-line. In addition to that, it provides with the flexibility of using more complex adaptation techniques viz. MLLR without increasing the overall latency. Moreover, one set of domain data adapted model means $\mathbf{m}^{r,(\alpha_i)}$ per α_i is required to be stored. As the range of α_i is limited to 0.88-0.98 in steps of 0.02, the overall memory requirement remains under check.

5.3.2.2 Global ML scaling of model mean parameters

As explained in 5.3.2.1, the selected model mean vectors are domain data adapted. Hence, those mean vectors are not guaranteed to be optimal for the given test data. To address this issue, we explored the test data dependent global scaling of the selected model mean vectors. In the this approach, given the warp factor α_i, the adapted mean vector $\boldsymbol{\phi}^r$ for the r^{th} Gaussian is modeled as

$$\boldsymbol{\phi}^r = \eta\, \mathbf{m}^{r,(\alpha_i)} \tag{5.15}$$

where $\mathbf{m}^{r,(\alpha_i)}$ is the mean vector for the r^{th} Gaussian in the chosen model $\mathbf{\Lambda}^{(\alpha_i)}$ corresponding to α_i and η is the scaling factor to be estimated in ML sense. The given test data $\mathbf{O}$ (a series of L observation sequences) is transformed using $\widetilde{\mathcal{H}}^{(\alpha_i)}$ to derive $\widetilde{\mathbf{O}}$ before computing the scaling factor. Following the procedure outlined in the CAT technique, an estimate of the scaling factor

is given by

$$\hat{\eta} = \Big[\sum_{r=1}^{R}\Big(\sum_{l=1}^{L}\gamma^r(l)\Big)\Big(\mathbf{m}^{r,(\alpha_i)}\Big)^T(\widetilde{\mathbf{C}}^{r,(\alpha_i)})^{-1}\mathbf{m}^{r,(\alpha_i)}\Big]^{-1}\Big[\sum_{r=1}^{R}\Big(\mathbf{m}^{r,(\alpha_i)}\Big)^T(\widetilde{\mathbf{C}}^{r,(\alpha_i)})^{-1}\Big(\sum_{l=1}^{R}\gamma^r(l)\,\tilde{\mathbf{o}}_l\Big)\Big] \tag{5.16}$$

where the $\gamma^r(l)$ *posteriori* probability of occupying the r^{th} Gaussian at time l given that the transformed observation sequence $\tilde{\mathbf{o}}_l$ is generated and $\widetilde{\mathbf{C}}^{r,(\alpha_i)}$ is the covariance matrix for the r^{th} Gaussian component in the chosen model $\mathbf{\Lambda}^{(\alpha_i)}$. The proposed fast adaptation approach is summarized in Algorithm 4.

Algorithm 4 SW-based Fast Adaptation Technique

Off-line part

Given: SI model $\mathbf{\Lambda}$, SW-based projection matrices $\widetilde{\mathcal{H}}^{(\alpha_i)}$ corresponding to the VTLN warp factors $\{\alpha_i\}_{i=1}^{6}$ lying in the range 0.88-0.98 in steps of 0.02, the domain-specific data from the child speakers and its true transcription.

for each α_i **do**

Step 1. Transform the means and the covariances in $\mathbf{\Lambda}$ using $\widetilde{\mathcal{H}}^{(\alpha_i)}$. Update all the parameters while deriving the lower dimensional model $\mathbf{\Lambda}^{(\alpha_i)}$.

Step 2. Modify the mean vectors in $\mathbf{\Lambda}^{(\alpha_i)}$ using the EV- or the MLLR-based adaptation and the correspondingly transformed domain-specific data to get $\mathbf{m}^{(\alpha_i)}$.

Step 3. Replace the means in $\mathbf{\Lambda}^{(\alpha_i)}$ by $\mathbf{m}^{(\alpha_i)}$ to get $\tilde{\mathbf{\Lambda}}^{(\alpha_i)}$.

end for

Decoding part

Step 4. Decode the test data $\mathbf{O} = \{\mathbf{o}_r\}_{r=1}^{R}$ using $\mathbf{\Lambda}$ to get the first-pass hypothesis.

Step 5. Compute the warped features $\mathbf{O}^{(\alpha_i)}$ corresponding to α_i

Step 6. Force-align each of $\mathbf{O}^{(\alpha_i)}$ with respect to $\mathbf{\Lambda}$ under the constraints of the first-pass hypothesis to compute its likelihood.

Step 7. Choose the value of α_i resulting in the highest likelihood.

Step 8. Select the model $\tilde{\mathbf{\Lambda}}^{(\alpha_i)}$ corresponding to the optimal α_i.

Step 9. Transform $\mathbf{O}$ using $\widetilde{\mathcal{H}}^{(\alpha_i)}$ to get $\widetilde{\mathbf{O}}$.

Step 10. Estimate global scaling factor $\hat{\eta}$ using equation in (5.16).

Step 11. Transform $\mathbf{O}^{(\alpha_i)}$ using $\widetilde{\mathcal{H}}^{(\alpha_i)}$ to get $\widetilde{\mathbf{O}}^{(\alpha_i)}$ and re-decode using $\tilde{\mathbf{\Lambda}}^{(\alpha_i)}$ with the r^{th} Gaussian mean vector replaced by

$$\boldsymbol{\phi}^r = \hat{\eta}\,\mathbf{m}^{r,(\alpha_i)}$$

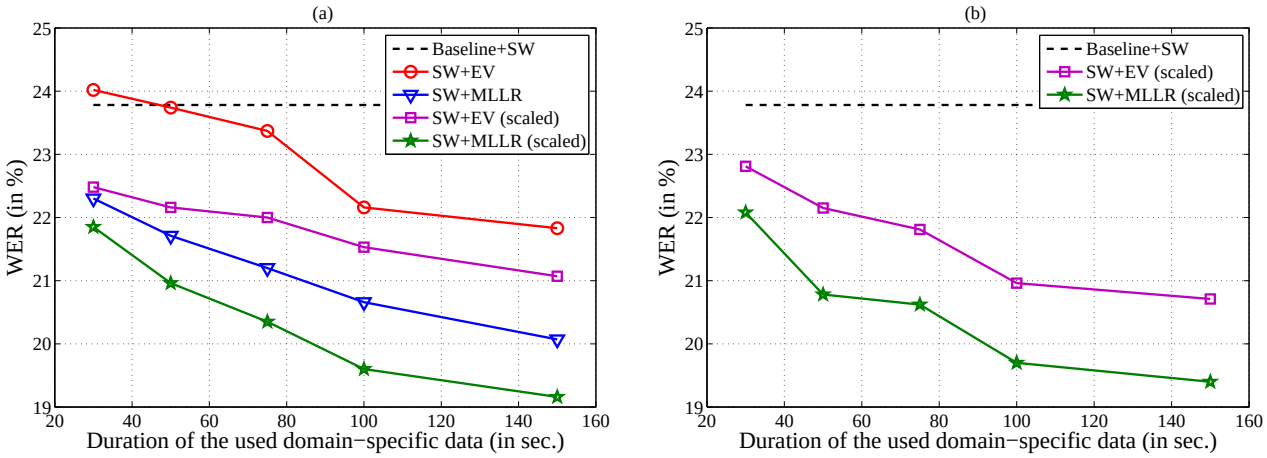

Figure 5.7: The WERs for the SW-based fast adaptation scheme in the (a) *utterance-specific* and (b) *incremental* modes. The x-axis represents the amount of children's domain-specific data used for the off-line estimation of model parameters.

5.3.2.3 Evaluation results

The recognition performances for the proposed approach in the utterance-specific mode with variations in the amount of domain-specific children's data used in deriving the pre-defined models are plotted in Figure 5.7(a). Recognition performances are shown for the cases when the mean vectors to be scaled are derived using the EV or the MLLR adaptation technique. Also shown are the WERs for the cases when no test data dependent scaling of the chosen mean vectors is performed. All the presented WERs include VTLN-based warping being applied to the test data. The effectiveness of the global scaling is quite evident from the shown recognition performances. The WERs for the same study performed in the incremental mode are shown in Figure 5.7(b). Similar trends are noted in this mode as well.

5.4 Summary

Motivated by the earlier works addressing the pitch-dependent distortions using a BW scheme, a structured low-rank feature projection technique has been proposed in this chapter. The use of low-rank projections also address the loss of relevant spectral information unlike the BW approach. As a consequence of applying the low-rank projections, the pitch dependent distortions are more effectively suppressed in the case of mismatched children's ASR. The SW-based projection is shown to result in a relative improvement of 35% over the baseline. Moreover, a novel fast adaptation approach for the GMM-HMM-based ASR system is also proposed. The presented technique exploits the correlation between the acoustic mismatch and

the estimated VTLN warp factor for the given test data. Based on the warp factor estimate, an appropriately transformed model is selected during testing. The selected model is then optimally scaled with respect to the test data. Consequently, the proposed approach involves the estimation of only two parameters. Despite these constraints, significant improvements are noted. The proposed adaptation approach is found to result in a relative improvement of 44% over the baseline. On account of a few parameter estimation, the overall computational cost is quite low. Moreover, it also ensures that the memory requirement is small enough. These factors make the proposed adaptation approach amenable for interactive ASR tasks.

An alternate way to address the effects of pitch-induced distortions is to provide sufficient smoothening of pitch harmonics during the acoustic feature extraction process. In the next chapter, we explore an existing approach that involves pitch-adaptive signal processing to obtain smoother spectra prior to the extraction of the final cepstral features. In addition to that, we also present a simple scheme for achieving the same.

6

Adaptive-Liftering-based Pitch Robust MFCC Features

Contents

In Chapter 5, we had presented a soft-weighting technique to suppress the variance of the higher-order cepstral coefficients. This approach was found to be very effective. The SW-based projections were applied on the trained acoustic models and the cepstral features of the test data. An alternate way of addressing this issue is to project the training data to a lower-dimensional subspace before training the acoustic models. This alternate way will be explored in Chapter 7. In this chapter, we explore the inclusion of the pitch-adaptive signal processing techniques in the computation of MFCC features. The use of these techniques is intended towards deriving the pitch-normalized MFCC features which would be relatively free from the aforementioned pitch-dependent distortions noted in the case of high-pitched signals.

One such technique is the STRAIGHT-based spectral analysis reported in [103]. In this approach, a pitch-adaptive window is employed during speech analysis which provides equivalent resolution in both time and frequency. The use of STRAIGHT-based MFCC features have already been explored in the context of adults' matched ASR [104]. In that study, the authors did not find the STRAIGHT-based MFCC features any better than the conventional MFCC features. In our understanding, such an observation is attributed to the ASR task being the matched one. The differences in the pitch across adult speakers is not as large as those existing between adult and child speakers. Thus, with the hope that the STRAIGHT-based MFCC features may turn out to be effective for children's mismatched ASR, it is being explored in this chapter. Further, we also present a simple scheme which adaptively discards the pitch information prior to computing the MFCC features. The proposed approach is found to be less sensitive to the errors in the estimation of the frame-specific pitch of the analyzed signal unlike the STRAIGHT-based approach. Both the explored pitch-adaptive feature extraction approaches are found to enhance the recognition performance of the mismatched ASR system. As already mentioned earlier, children's speech has higher formant frequencies and greater spectral variability [22, 23]. Consequently, we have also explored the feature-space normalization approaches in combination with the pitch-adaptive MFCC features. The combination of the techniques is found to be highly effective in the case of the mismatched ASR task.

The remaining of this chapter is organized as follows: In Section 6.1, the STRAIGHT-based spectral analysis and feature extraction is discussed. In Section 6.2, the proposed adaptive-liftering-based pitch-robust feature extraction approach is outlined. The impact of the explored pitch-adaptive features are empirically analyzed in Section 6.3. The experimental evaluation of the pitch-adaptive schemes is discussed in Section 6.4. The effect of combining the pitch-adaptive features with the VTLN and the SW is presented in Section 6.5 Finally the chapter

is concluded in Section 6.6.

6.1 Review of STRAIGHT-based MFCC Features

In contrast to the conventional STFT-based spectral analysis employing fixed window, the STRAIGHT spectral analysis involves a pitch-adaptive window having equivalent resolution in both time and frequency domains [103]. In the STRAIGHT analysis, the speech signal is windowed with two complimentary pitch-adaptive windows $w_p(t)$ and $w_c(t)$. These windows are based on the product of a Gaussian function and a 2^{nd}-order cardinal B-spline function, and those are given by

$$w_p(t) = e^{-\pi(t/T)^2} \circledast b(t/T) \tag{6.1}$$

$$w_c(t) = w_p(t)\sin(\pi\frac{t}{T}) \tag{6.2}$$

where t is the time index, T is the pitch (fundamental) period ($T = 1/F_0$) and $b(t)$ is the Bartlett window given by

$$b(t) = \begin{cases} 1 - |t|, & |t| < 1 \\ 0, & \text{otherwise} \end{cases}. \tag{6.3}$$

The final smoothed spectrum is obtained as

$$P(\omega, t) = \sqrt{P_p^2(\omega, t) + \zeta P_c^2(\omega, t)} \tag{6.4}$$

where $P_p^2(\omega, t)$ and $P_c^2(\omega, t)$ are the power spectrum obtained using $w_p(t)$ and $w_c(t)$, respectively. The blending factor value $\zeta = 0.13655$, as suggested in [103], is used. The partial information so obtained is interpolated to derive a smoothed time-frequency representation that is devoid of the ill effects caused by the signal periodicity. This spectral analysis has been effectively employed in speech synthesis [105, 106] and voice conversion [107, 108].

For the STRAIGHT-based spectral analysis, a reliable estimate of the pitch is very important. Errors in the pitch estimation lead to improper cancellation of pitch harmonics, thus resulting in spectral distortions. For this reason, the pitch is estimated using the TEMPO algorithm [103] which is noted to have a high latency. The MFCC features derived using the STRAIGHT-based spectral analysis were employed for adults' ASR in [104] and were found to be inferior to the conventional ones. Later the authors analyzed the cause of the degradation in a smoothening function used after the pitch-adaptive windowing in the STRAIGHT processing. On removing that smoothening function, an enhanced recognition performance was obtained as reported in [109]. Following those studies, we explore the effectiveness of the STRAIGHT-

based MFFC features in the context of children's mismatched ASR. Since the pitch-dependent distortions are severe in those cases, large improvements in the recognition performance are expected.

6.2 Adaptive-Liftering-based MFCC Features

The need of robust frame-wise pitch estimate makes the STRAIGHT approach less amenable for applications involving human-machine interaction. The high sensitivity of STRAIGHT approach to the accuracy of the pitch estimates is already pointed out in [104]. In fact, when the STRAIGHT-based MFCC features are computed using a global pitch value (i.e., averaged over all voiced frames in an utterance) rather than the frame-specific pitch values, the corresponding recognition performance is noted to fall below that obtained for the conventional MFCC features. Motivated by these issues, we present an alternate scheme for smoothing the spectrum exploiting adaptive cepstral liftering. Being based on much simpler premise, it does not require an accurate frame-wise pitch estimate to be effective.

The steps in the proposed scheme are as follows: First the spectral representation of the speech signal is obtained using the STFT analysis with fixed duration Hamming window. For each frame, the log-compressed magnitude spectrum is derived and is transformed to cesptral domain using an inverse discrete Fourier transform (IDFT). Note that all these processing steps are essentially equivalent to linear filtering, thus the periodicity of the speech excitation is retained in the cepstral domain. Now a suitable low-time lifter window is applied and the liftered cepstrum is transformed back to the spectral domain using the DFT. The block diagram of the proposed smoothed spectrum derivation approach is shown in Figure 6.1. Given the smoothed spectrum, the MFCC features are derived following the usual steps outlined in Appendix A.

The duration of the low-time lifter is chosen based on the average pitch value for the utterance being analyzed. In this work, a cepstral-domain-based pitch detection algorithm, outlined in [110], is used for the estimation of the pitch. For checking the consistency, we also performed the pitch estimation using a few other algorithms, viz. TEMPO [103], RAPT [111] and WaveSurfer [112]. On comparing, some differences in the frame-specific pitch estimates were noted among these algorithms, but their average pitch values turned out to be close enough. Also a liftering window with slanting right-edge is used to avoid ripples in the derived smoothed spectrum.

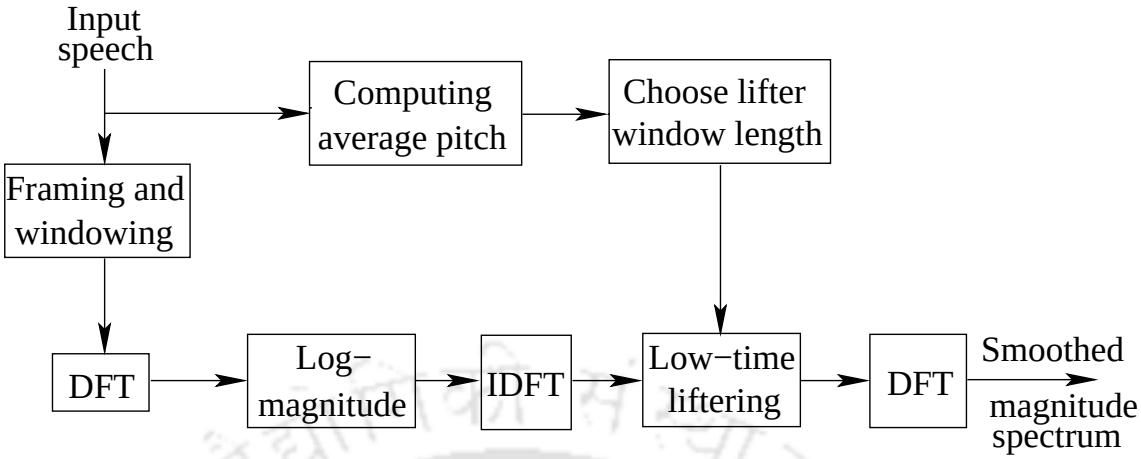

Figure 6.1: Block diagram of the proposed pitch-adaptive liftering approach for spectral smoothing.

6.3 Analyzing the Impact of Pitch-Adaptive Signal Processing

In Chapter 4, the effect of pitch-induced acoustic mismatch was analyzed. It was noted that, due to the pitch-dependent distortions in the case of high-pitched speakers, some ripples in the spectral envelope appeared particularly in the lower-frequency region of the speech spectrum. The primary motivation behind employing the pitch-adaptive MFCC features is to smoothen out the ripples in the envelope of the obtained spectrum. The log-compressed magnitude spectra obtained by the conventional approach employing static signal processing for two different vowels are shown in Figure 6.2. The degree of spectral smoothening achieved through the proposed adaptive-liftering-based approach with variations in the length of the applied lifter window is also shown in Figure 6.2.

Another way to visualize the degree of spectral smoothening is by transforming the cepstra back to the spectral domain and plot the same. Figure 6.3 depicts the relative degree of smoothening obtained through the two explored pitch-adaptive approaches. Compared to the spectrum corresponding to the conventional/static MFCC features (shown in Figure 4.3), the ripples in the low frequency region are effectively removed by both the techniques. It is to note that a much smoother spectrum is obtained by the STRAIGHT-based approach. In the case of the proposed technique, some ripples are still present in the low frequency regions especially for the high pitch case, $F_0 = 300$ Hz. This is due to the fact that a lifter of fixed duration computed globally is applied to all the frames. The change in the variance of the higher-order coefficients due to the pitch-adaptive analysis is shown in Figure 6.4. Compared to the static

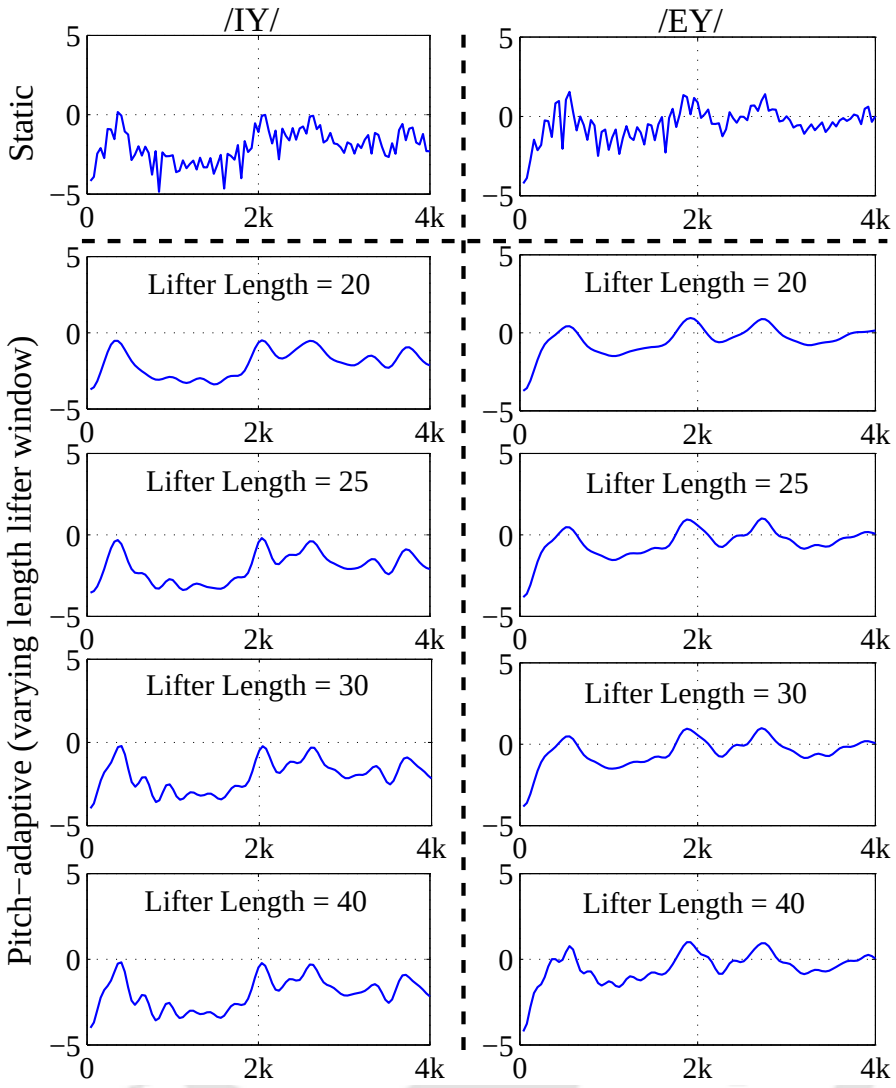

Figure 6.2: Demonstration of the spectral smoothing effected by the proposed approach. The left and the right panels show the log-compressed magnitude spectra for a high-pitched (F_0 = 300 Hz) speech frame for the vowels /IY/ and /EY/, respectively, collected from TIMIT. In these panels, the x-axis denotes the frequency values in Hz.

MFCC features, the mismatch in the variances is reduced significantly. Similar trends are also noted for the case of children's test set as shown in Figure 6.5. This reduction in the mismatch in the variance is expected to improve recognition performances in the mismatched ASR cases.

6.4 Experimental Evaluation

6.4.1 ASR system specifications

For extracting the static MFFC features, a Hamming window of length 25 ms with a frame shift of 10 ms is employed for speech data analysis. Employing a 21-channel Mel-filterbank, first

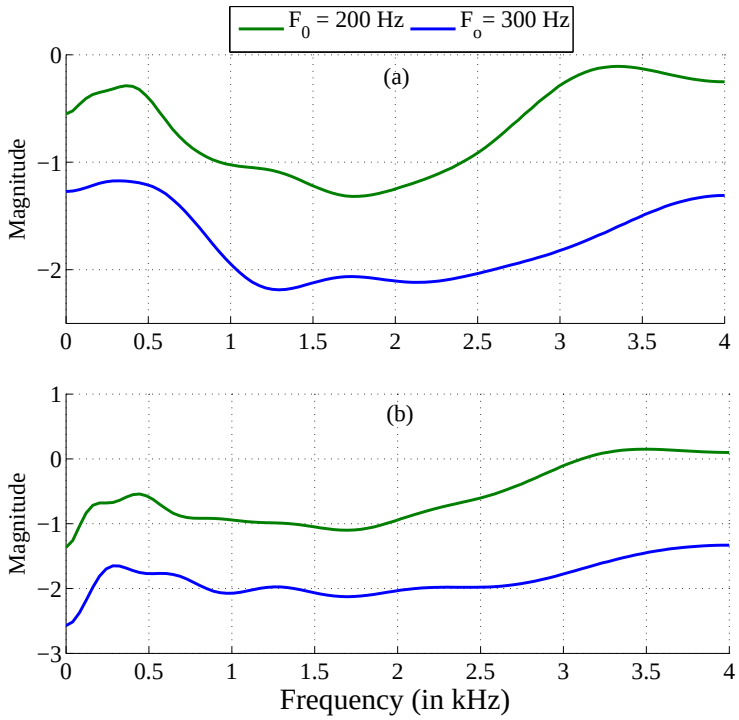

Figure 6.3: The smoothed spectrum obtained using a the two techniques (a) STRAIGHT (b) proposed, are shown for the cases when F_0 = 200 Hz and F_0 = 300 Hz. In the case of the proposed approach, the duration of the applied lifter is determined using the average pitch and the optimally smoothed spectrum is plotted. To make the curves distinguishable, an intentional shift of 2 dB is added.

12-dimensional MFCC features are computed (C_1-C_{12}). The log power is added as the zeroth coefficient making the base feature dimension equal to 13 (C_0-C_{12}). Two different post-processing approaches are employed on top of base MFCC features to derive the final feature vectors used in acoustic modelling. In the first approach, the base MFCC features are augmented with their first- and second-order temporal derivatives (computed over a span of 5 frames) to yield 39-dimensional feature vectors. These features are referred to as the TYPE-I features. In the second approach, the base MFCC features are spliced in time considering a context of 4 frames on either side of the current frame. The dimensionality of the obtained 117-dimensional vector is reduced to 39 through linear discriminant analysis (LDA) [113]. The resulting low dimensional vector is further decorrelated using maximum likelihood linear transform (MLLT) [114] (also referred to as global semi-tied covariance (STC) [115]). The latter approach is reported to yield better recognition performance and the derived features are referred to as the TYPE-II features.

In the case of the STRAIGHT-based analysis, the pitch estimation is done frame-wise. In

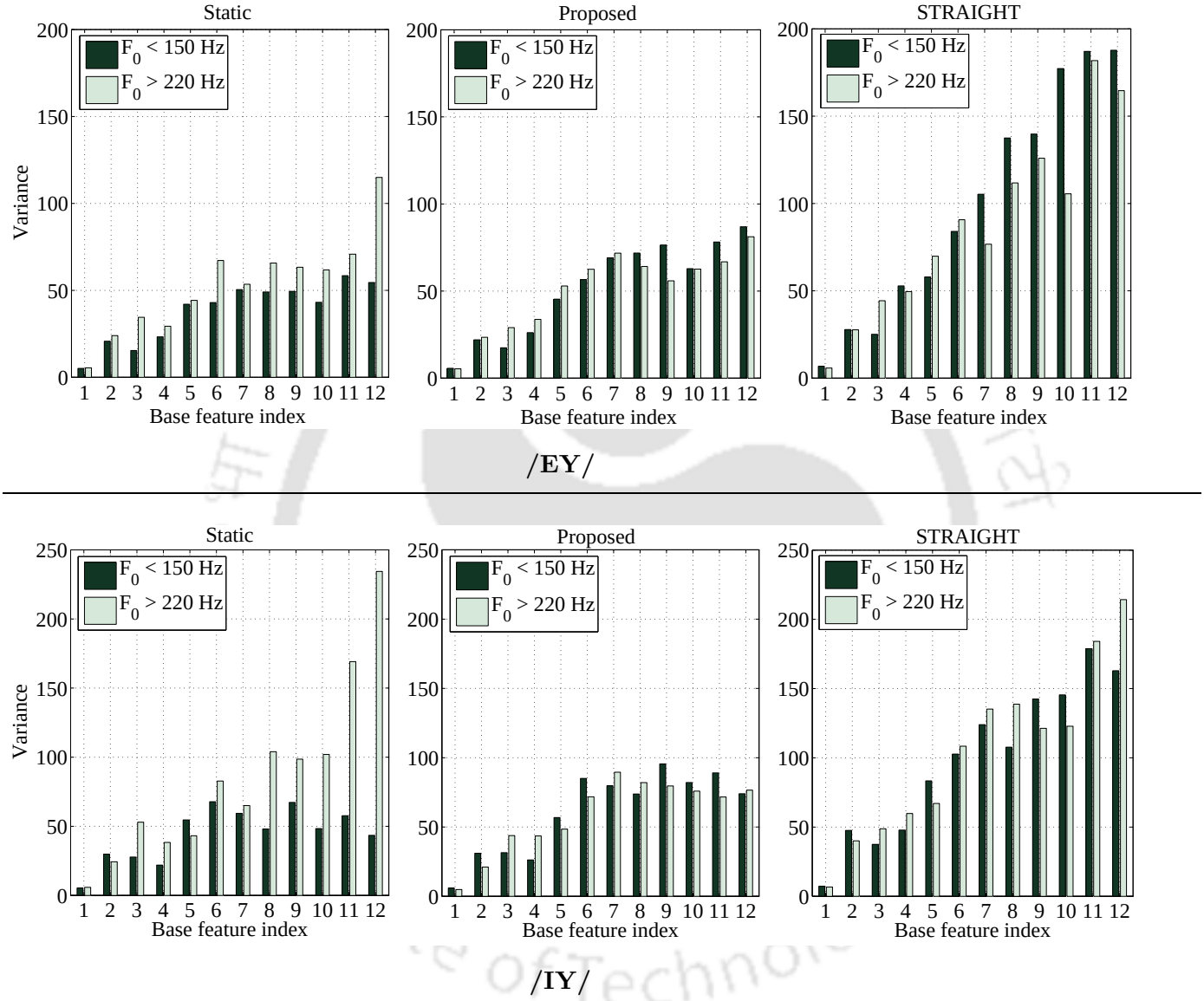

Figure 6.4: Variance of the base MFCC features (C_1-C_{12}) for the vowel /EY/ (top panel) and /IY/ (bottom panel) for two broad pitch (F_0) ranges. Figure also shows the reduction in the variance mismatch as a result of the pitch harmonic smoothening achieved by the proposed and the STRAIGHT based approach.

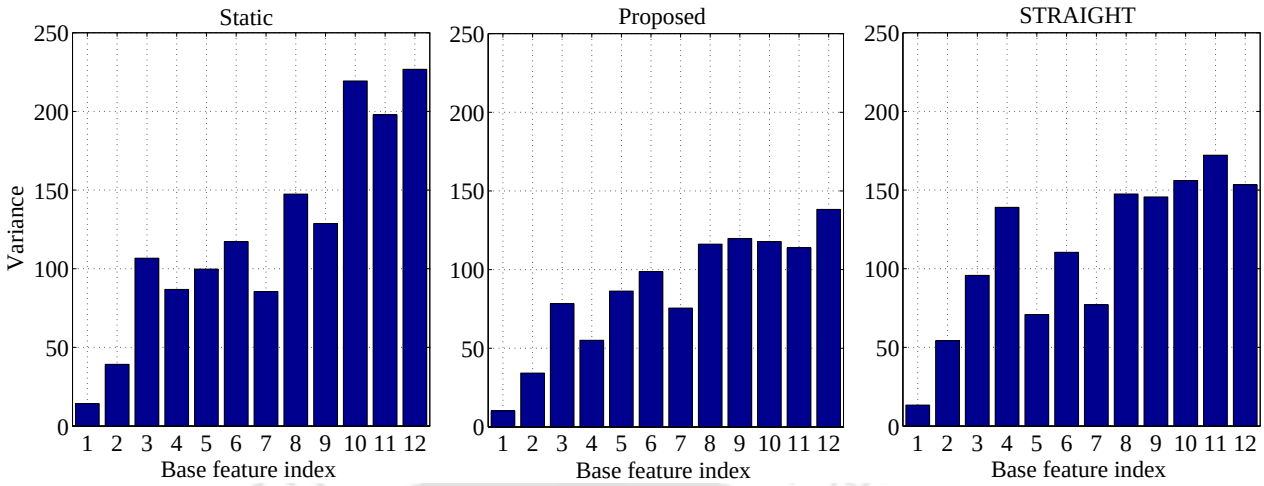

Figure 6.5: Variance of the base MFCC features (C_1-C_{12}) for the vowel /IY/ for the children's data. Note the decrease in variance due the spectral smoothening achieved by the pitch-adaptive approaches.

the default STRAIGHT analysis, it is recommended to use a frame length of 80 ms and a frame shift of 1 ms. This would lead to very high latency as well as the frame rate would differ from that used for the proposed adaptive-liftering-based approach. Therefore, we have kept a frame length of 25 ms along with a frame shift of 10 ms for the STRAIGHT-based and the adaptive-liftering-based MFCC feature computation. For the conventional as well the proposed pitch-adaptive MFCC feature extraction, the VOICEBOX [116] (a MATLAB toolbox), is used making appropriate modifications when needed. Time-splicing followed by LDA and MLLT (TYPE-II) is explored in the case of pitch-adaptive features as well. The ASR system employed in the studies reported in this chapter is developed using the Kaldi toolkit [10, 117]. The other details of the employed experimental setup remain the same as given in Section 3.3.

6.4.2 Evaluation results

The word error rates for the two kinds of static features (TYPE-I & II) are given in Table 6.1. It is to note that the WER obtained for the static TYPE-I features is quite similar to those reported in the previous chapters. This shows that the features extracted using the HTK and the MATLAB toolboxes have resulted in very similar recognition performances. Moreover, the effect of employing the Kaldi and the HTK toolkits in developing the ASR system is also noted to be minimal. The WERs for the pitch-adaptive features explored in this work are also given in Table 6.1. Separate ASR systems are trained using the three kinds of the MFCC features. The evaluation of the mismatch recognition performance for a particular kind of MFCC features is done on an ASR system developed using the matching feature kind. The

Table 6.1: The WERs for the three explored feature extraction approaches with respect to the ASR systems trained using the adults' speech. The two discussed post-processing approaches are applied to the base MFCC features and the WERs are given for both the cases.

MFCC kind	WER (in %)	
	Type-I	Type-II
Default (Static)	64.13	62.55
STRAIGHT-based	55.24	52.34
Adaptive-liftering-based	54.06	50.78

Table 6.2: The WERs on adults' matched (CAMts) and children's mismatched (PFts) test sets for the explored variants of MFCC features with respect to GMM-HMM- based ASR systems trained using the adults' speech data.

MFCC kind	WER (in %)	
	CAMts	PFts
Default (Static)	12.15	62.55
STRAIGHT-based	11.21	52.34
Adaptive-liftering-based	10.97	50.78

use of the TYPE-II features results in a slight improvement in the recognition performance. Consequently, the TYPE-II post-processing is followed in the remaining of this thesis and is referred to the conventional/static features. The use of pitch-adaptive MFCC features results in much better recognition performances than that of the static features. Furthermore, both the explored pitch-adaptive approaches result in quite similar recognition performances. The effectiveness of using the pitch-adaptive MFCC features (TYPE-II) in the matched as well as the mismatched cases are given in Table 6.2. Note that both the pitch-adaptive approaches result in similar recognition performances in the matched case as well.

6.5 Combining with Existing Mismatch Reduction Techniques

The effect of combining some of the existing techniques like the VTLN, the fMLLR and the low-rank feature projection (presented in the previous chapter) with the proposed pitch-adaptive features is evaluated in the following subsections.

Table 6.3: The WERs for the conventional/static and the pitch-adaptive feature extraction approaches in combination with the VTLN- and the fMLLR-based feature normalization. It is to note that the 95% confidence intervals for the performance with respect to the fMLLR and the VTLN included baselines are ±1.37 and ±1.30, respectively. Hence, the observed improvements in the recognition performances are statistically significant.

MFCC kind	WER (in %)	
	VTLN	fMLLR
Default (Static)	35.06	43.53
Adaptive-liftering-based	27.62	35.83

6.5.1 Vocal tract length normalization

In the previous chapters it was noted that the VTLN was quite powerful in the case of the children's mismatched ASR. In order to implement VTLN, the spectrally warped MFCC features corresponding to different values of the linear warp factors are derived. Again, the warp factor values chosen in this work lie in the range of 0.88-1.12 in steps of 0.02. The ML grid search is performed to choose the optimal warp factor value for a test utterance. Optimally warped spectral features are then decoded with respect to the baseline system to generate an enhanced hypothesis. The effect of applying the VTLN-based frequency warping on the default and the proposed cepstral features is given in Table 6.3. Significant improvements in the WERs are achieved when spectral warping is performed on both the static as well as the pitch-adaptive MFCC features. Moreover, the overall WERs still happen to be better in the case of the pitch-adaptive cepstral features.

6.5.2 Feature-space maximum likelihood linear regression

Another feature normalization approach reported is the feature-space MLLR (fMLLR) or the CMLLR. In order to implement the fMLLR, the required transform is estimated using the SI system applying speaker adaptive training (SAT) [51, 118]. The derived fMLLR transform consists of 39×40 parameters. During testing, a fMLLR transform is estimated for the given test utterance and then applied in decoding. The WERs for this study are also given in Table 6.3. Though the fMLLR-based feature normalization is also noted to be effective similar to the VTLN, but the computational cost of the former happens to be much higher than that noted for the later.

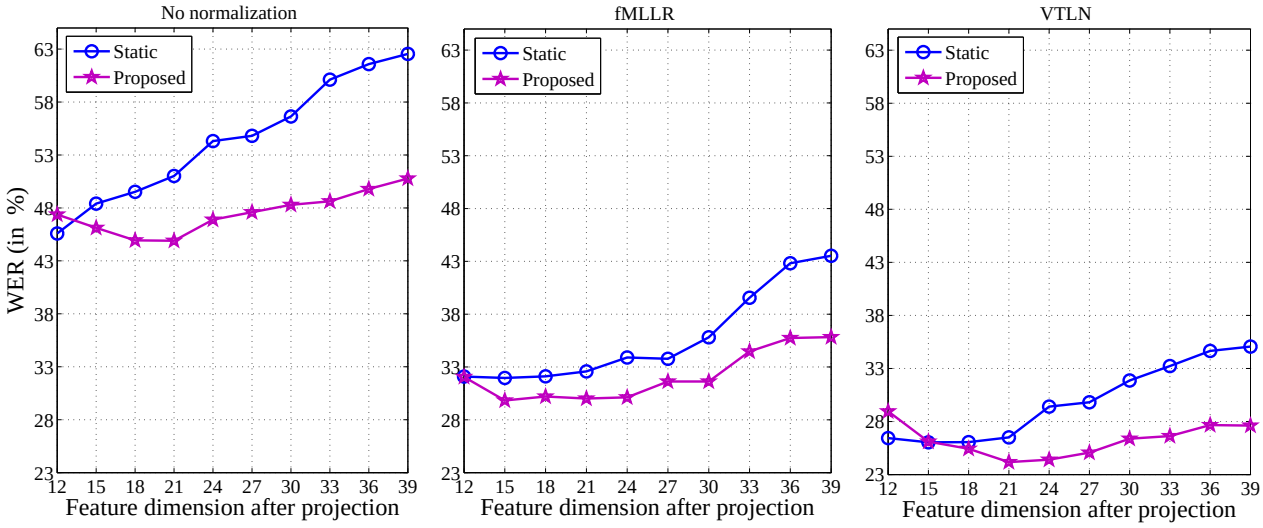

Figure 6.6: Recognition performances for the SW-based projection applied to the proposed pitch-adaptive cepstral features with and without feature normalization.

6.5.3 Structured low-rank feature projection

The SW-based projection is employed to the default as well as the proposed pitch-adaptive cepstral features along with the earlier discussed feature normalization techniques. The WER-profiles with varying ranks of the projection matrix are shown in Fig 6.6. Since the 95% confidence intervals for the performance with respect to the fMLLR and the VTLN included baselines are ± 1.37 and ± 1.30, respectively, the observed changes are statistically significant. It is quite evident that employing the SW on the proposed features leads to better recognition performances. Further, a lesser suppression of the higher dimensions is required since the pitch-dependent distortions have been significantly suppressed during the feature extraction phase itself.

6.6 Summary

In this chapter, the role of pitch-adaptive feature extraction processes in the context of children's mismatched ASR has been explored. In this regards, the existing STRAIGHT-based pitch-adaptive signal processing is employed during speech parameterization. Furthermore, we have also presented a simple pitch-adaptive scheme for feature extraction. The proposed approach happens to be less sensitive to errors in the estimation of the pitch than the STRAIGHT. The pitch-adaptive feature computation approaches tend to smoothen out the pitch-dependent distortion thus reducing the aforementioned variance mismatch. The same has been verified

experimentally in this chapter.

All the experimental studies presented this far involved the GMM-HMM-based ASR systems. In the last few years there has been a paradigm shift in the speech recognition research. The GMMs are now being replaced by the DNNs. Consequently, we have explored the effectiveness of the approaches proposed in Chapters 5 and 6 on DNN-based ASR systems in Chapter 7. We also explore another acoustic modelling technique based on the SGMM for the sake of completeness. The ASR systems employing the SGMM-based acoustic modelling are generally reported to be better than those based on the GMM.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

7

Exploring Pitch Compensation in SGMM and DNN Domains

Contents

Recently developed acoustic modelling techniques based on the subspace Gaussian mixture model (SGMM) and the deep neural network (DNN) have been reported to achieve significant improvements in the matched case recognition performance over the conventional GMM-based modelling. Motivated by that, in this chapter, we explore those acoustic modelling techniques in the context of the children's mismatched ASR. To the best of our knowledge, only a few works on children's ASR employing DNN-based acoustic modelling have been reported [119–121]. Further, the proposed low-rank feature projection and pitch-adaptive cepstral features are also explored in the context of those modelling approaches.

The remaining of this chapter is organized as follows: In Section 7.1, the details of the setup employed in the development of the SGMM- and the DNN-based ASR systems are discussed. A discussion on the baseline recognition performances for the developed systems in the context of children's mismatched ASR task is presented in Section 7.2. In Section 7.3, we study the effect of employing the low-rank feature projection in the case of SGMM- and DNN-based systems. Next, the role of adaptive-liftering-based acoustic features is studied in Section 7.4. Finally, the chapter is concluded in Section 7.5.

7.1 Experimental Setup

In this section, we first provide the details of the employed experimental setup. Since the used speech corpora are the same as that in the earlier chapters, only the new inclusions are being discussed. The adults' speech trained GMM-, SGMM- and DNN-based ASR systems employed for experimental evaluation are developed using the Kaldi toolkit [10, 117]. For all the explored acoustic modelling approaches, the 39-dimensional TYPE-II static features are employed. In order to further improve the recognition performance, the default features are normalized using fMLLR as proposed in [122]. The required fMLLR transform is estimated using the GMM-HMM-based system applying speaker adaptive training (SAT). The said normalization is applied to both training and test features.

For the GMM-HMM system development, the cross-word tri-phone acoustic model training along with the decision tree-based state tying is employed for the same. Each context-dependent triphone (senone) is modeled using a 3-states HMM with 8 diagonal covariance Gaussian components per state. The number of senones in the GMM-HMM-based system is kept as 2000. For the SGMM training, the number of Gaussians used in the training of the universal background model (UBM) is selected as 400. The number of leaves and Gaussians in the SGMM are chosen to be 9000 and 7000, respectively. In our experimental setup, the subspace dimension is kept

same as the feature dimension in the SGMM-based training (may refer to Appendix B.2 for details). These parameters are found to be suitable for the matched case testing. In the case of GMM-HMM as well as the SGMM-HMM, separate ASR systems are trained using the default (unnormalized TYPE-II) and the normalized features.

In the case of DNN-HMM-based system, the nonlinearities in the hidden layers is the *tanh* function. The employed objective function is the cross-entropy criterion, i.e., the log-probability of the correct class for each of the frames. An initial learning rate of 0.015 is selected which is reduced to 0.002 in 20 epochs. Extra 10 epochs are employed after reducing the learning rate to 0.002. In Kaldi, a preconditioned form of stochastic gradient descent is employed during the DNN training. In this approach, instead of using a scalar learning rate, a matrix-valued learning rate is used. This is motivated by the basic idea to reduce the learning rate in dimensions where the derivatives have a high variance. This approach, in turn, is to control instability and stop the parameters moving too fast in any one direction. The minibatch size for neural net training is selected as 512. The number of hidden layers is varied from 2 to 8 and finally fixed at 8. In Section 7.3.1, we present a brief discussion on the effect of varying the number of hidden layers in the context children's mismatched ASR. The output layer is a soft-max layer and the outputs represent the log-posterior probability of the output labels corresponding to context-dependent HMM states. Separate DNNs are learned on the default (unnormalized TYPE-II) and the fMLLR-normalized features. For the default-feature-based DNN training, the decision tree and the state alignments required as the supervision are obtained from the corresponding GMM-HMM system. In the case of fMLLR-based DNN, a revised GMM-HMM system with same complexity is developed using the normalized features. The decision tree and the state alignments are obtained from the revised system.

7.2 Baseline System Evaluation

The recognition performances of the baseline ASR systems developed using the three kinds of acoustic modelling techniques with respect to the two earlier mentioned test sets are given in Table 7.1. For both the test sets, the SGMM- and the DNN-based acoustic modellings are found to be superior to the GMM-based one as expected. Further, both the SGMM- and the DNN-based systems exhibit a large degradation in the mismatched testing case similar to that noted for the GMM-based one. It is to be noted that the WERs are being given for the fMLLR-normalized features as those are reported to be much superior to the default features in the case of matched ASR. Since the focus of this work is on children's mismatched ASR, we

Table 7.1: The WERs for adults' speech trained SI system under acoustically matched and mismatched test conditions with different acoustic modelling approaches. All the reported performances include LDA, MLLT and fMLLR-based transformations being applied to features. The 95% confidence intervals for the performance (PFts) with respect to the fMLLR included baselines for GMM-, SGMM- and DNN-based systems are ±1.37, ±1.28 and ±1.20, respectively.

Speech data used for SI system training	Acoustic modelling approach	WER (in %)		% Relative difference w.r.t. mismatched
		Matched (CAMts)	Mismatched (PFts)	
CAMtr	GMM-HMM	8.13	43.53	81
	SGMM-HMM	6.81	31.96	79
	DNN-HMM	6.20	24.25	73

Table 7.2: The WERs for the children's test on ASR systems employing different kinds of acoustic models with and without the fMLLR/VTLN. Note that the 95% confidence intervals for the performance (PFts) with respect to the VTLN included baselines for GMM-, SGMM- and DNN-based systems are ±1.30, ±1.21 and ±1.21, respectively.

Speech data used for SI system training	Acoustic model kind	WER (in %)		
		Default	fMLLR	VTLN
CAMtr	GMM-HMM	62.55	43.53	35.06
	SGMM-HMM	54.43	31.96	26.12
	DNN-HMM	43.32	24.25	23.57

will be discussing the effect of normalization in detail in the following subsection.

7.2.1 Inclusion of feature-space normalization

In the previous chapters, it has been shown that the recognition performance of the children's mismatched ASR can be improved significantly by the use of VTLN-based frequency warping. In order to study the effect of VTLN in the SGMM/DNN domain, we explored frequency warping during the mismatch testing. All three (GMM, SGMM, and DNN) systems are developed using the default features (unnormalized TYPE-II) for this study. For the children's test set, the default features are warped prior to decoding to quantify the effect of the VTLN. The WERs for those experiments are given in Table 7.2. For better contrast, the fMLLR-normalized features are also explored. The correspondingly trained GMM-, SGMM- and DNN-based systems are employed for this study and the WERs are also given in Table 7.2. It can be noted that both the fMLLR and the VTLN are very effective in the mismatched

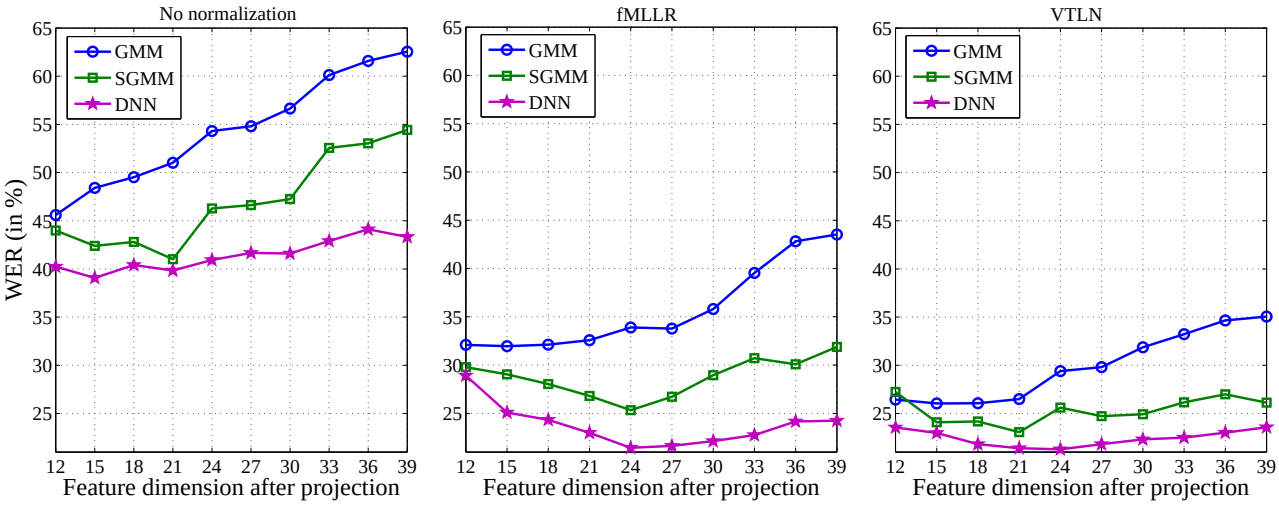

Figure 7.1: The WER profiles with variations in the rank of feature projection matrix for the three acoustic modelling approaches namely GMM, SGMM and DNN. Also shown are the effects of normalizing the time-spliced base MFCC features using either the fMLLR or the VTLN.

testing case. In our experiments, the optimal VTLN-based warping is implemented using a low-complexity ML grid search. On the other hand, the fMLLR-based normalization requires the estimation of 39×40 parameters, thus involving much higher computational cost.

7.3 Revisiting Low-Rank Feature Projection

From the studies presented in the previous sections, it can be concluded that both SGMM- and DNN-based ASR systems are more effective than GMM one even in the context of mismatched task. Normalization techniques lead to further reductions in the WERs in all the three cases. The relative differences in the mismatched and the matched case recognition performances are given in the last column of Table 7.1. It can be noted that, even after normalization, there is still a huge gap between the recognition performances for the adult and the child speakers. Consequently, for improving the recognition performance of children's speech on the SGMM- and the DNN-based systems, we have also explored low-rank feature projection. The architecture of these two acoustic modelling approaches is not the same as the GMM-based system. Therefore, the procedure outlined in Chapter 5 cannot be used as it is. An alternate means to do the same is to project the training speech features to a lower dimensional subspace before learning the acoustic models. For this purpose, the rank of the LDA projection matrix employed on the time-spliced base MFCC features is reduced. The acoustic models are then trained using those reduced dimensional features.

Table 7.3: Percentage relative improvement (PRI) in the recognition performances obtained through the use of low-rank feature projection. the WERs are reported for the three acoustic modelling approaches explored. Also shown are the WERs for the cases when the fMLLR/VTLN is applied for feature normalization. Note that the 95% confidence intervals for the performance (PFts) with respect to the VTLN+fMLLR included baselines for GMM-, SGMM- and DNN-based systems are ±1.20, ±1.10 and ±1.03, respectively.

Acoustic modelling approach	WER (in %)											
	No normalization			VTLN			fMLLR			VTLN + fMLLR		
	Def.	SW	PRI	Def.	SW	PRI	Def.	SW	PRI	Def.	SW	PRI
GMM	62.55	45.59	27.11	35.06	26.04	25.73	43.53	31.96	26.56	25.82	18.62	27.89
SGMM	54.43	41.01	24.66	26.12	23.06	11.72	31.89	25.33	20.57	19.76	17.22	12.86
DNN	43.32	**39.07**	9.81	23.57	**21.40**	9.21	24.25	**21.44**	11.58	17.00	**15.40**	10.58

The effect of the low-rank feature projection on children's mismatched ASR can be accessed by the WER profiles for the earlier discussed modelling approaches shown in Figure 7.1. The WER profiles are shown with and without the fMLLR being included in training and testing. The fMLLR-based normalization is found to result in additive reductions in WERs. The WER-profiles for the case when the VTLN is employed instead of the fMLLR are also shown for all three modelling cases. This is for assessing the relative advantage of the VTLN/fMLLR for the children's mismatch ASR. Both the fMLLR and the VTLN are observed to result in additive reductions in WERs when combined with SW, i.e., the low-rank projection. For highlighting the relative reduction in WERs for the different modelling approaches, the best case performances with low-rank projection are given in Table 7.3. It is clearly evident that the the low-rank feature projection results in significant improvements in the recognition performances in all the explored combinations. Moreover, it is to note that the combination of the fMLLR with the VTLN leads to further reductions in the WERs. But this reduction is achieved at an increased computational cost.

7.3.1 Effect of varying the number of hidden layers in DNN

For the matched testing case, a relative reduction of 22% in WER over the GMM-HMM system was obtained with the DNN-based acoustic modelling. Most of the earlier works had reported similar trends in the matched testing case. Varying the number of hidden layers did not result in much changes in the matched case recognition performance may be because the amount of training data was moderate. Consequently, we were confident that the developed

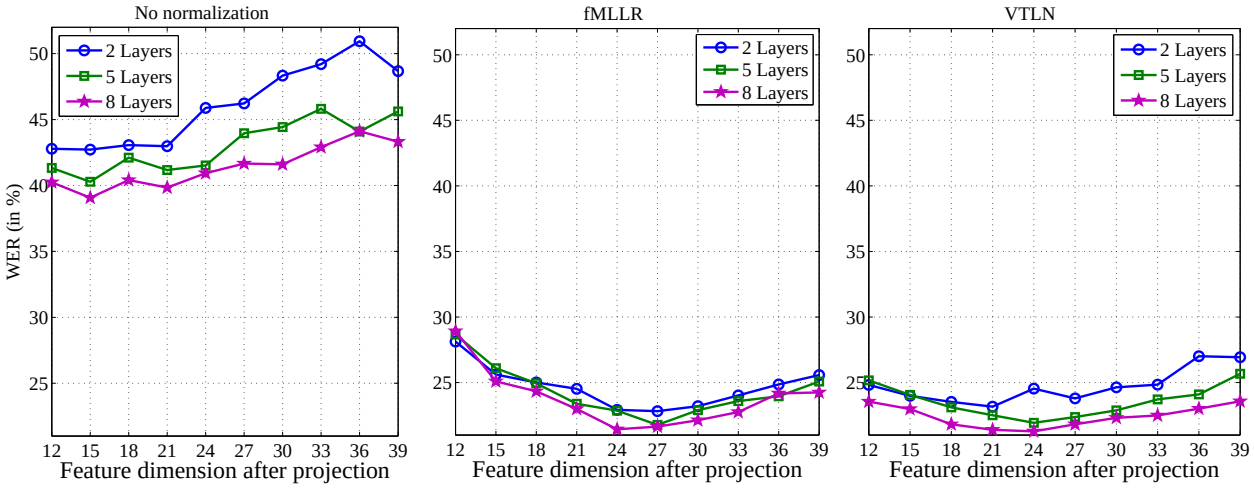

Figure 7.2: The WER profile with variations in the rank of feature projection matrix with respect to the DNN-HMM systems involving 2, 5 and 8 hidden layers, respectively. Also shown are the effects of normalizing the time-spliced base MFCC features using either the fMLLR or the VTLN.

DNN-HMM system involving 8 hidden layers was good enough and proceeded with our study in the mismatched case.

In the studies presented in Section 7.3, the VTLN was observed to be quite effective. On the contrary, the VTLN was found to be largely ineffective in the case of DNN-HMM systems with large number of hidden layers (7+) as reported in [123]. According to that study, the VTLN is found to be effective in the case of shallow networks only. Again, the experiments reported in that work are for the matched case testing only. Thus, it would be worth confirming whether the same trend holds in the case of mismatched testing explored in this work or not. Consequently, we varied the number of hidden layers for the mismatched testing going from shallow to deep networks. At the same time, the low-rank projection of the data was also explored in each of the cases.

The effect of projecting the data to a lower dimensional subspace on the three different complexity DNN-HMM systems is shown in Figure 7.2. The WER profiles are shown for the cases when the number of hidden layers is 2, 5 and 8 only to avoid overcrowding of the plots. Also shown are the WER profiles when the acoustic features are normalized using the fMLLR as well as the VTLN. It is interesting to note that the effect of varying the number of hidden layers is much more pronounced when no feature normalization is included. Moreover, the effectiveness of VTLN remains intact in all the studied cases. Moreover, the WERs obtained through the use of fMLLR and VTLN turn out be quite similar in all the cases. In addition to that, the low-rank feature projection is observed to result in further reductions in WERs for

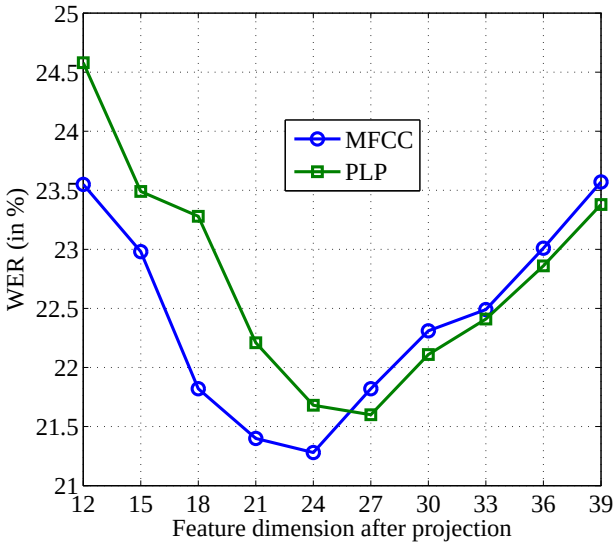

Figure 7.3: The WER profile with variations in the rank of feature projection matrix for the DNN-HMM system with respect to the PLP features. The reported WERs include VTLN-based warping of the test data.

the studied cases.

7.3.2 Experiments with PLP features

The use of perceptual linear predictive (PLP) features has been successfully explored in the conventional ASR systems [123]. As shown in Chapter 5, the low-rank projection was found to be effective in the case of PLP features as well. In this subsection, we explore the use of PLP features in the context of DNN system along with low-rank feature projection. The WER profile for this study on a DNN-HMM system is shown in Figure 7.3. The plotted WER profiles include the VTLN-based frequency warping of the test data. It is evident from the presented WERs that the use of low-rank projection is similarly effective for the PLP features as noted for the MFCC features in the context of DNN-based system.

7.4 Role of Pitch-Adaptive Cepstral Features

In the previous chapter, the pitch-adaptive MFCC features were shown to be quite effective for the children's mismatched ASR. In the following we revisit the same and study the effect of pitch-adaptive signal processing in the context of SGMM- and DNN-based ASR systems. This is followed by a study on the effect of feature normalization and low-rank projection on the pitch-adaptive MFCC features in the context of SGMM- and DNN-based systems.

Table 7.4: The WERs for the static and the pitch-adaptive MFCC feature extraction approaches explored in this work. The WERs are enlisted for the mismatched testing with respect to the ASR systems employing the GMM-, the SGMM- and the DNN-based acoustic modelling.

Feature kind	WER (in %)		
	GMM	SGMM	DNN
Default (Static)	62.55	54.43	43.32
STRAIGHT-based	52.34	49.36	39.32
Adaptive-liftering-based	50.78	48.35	39.30

The recognition performances for the three kinds of MFCC features explored in this work are given in Table 7.4. Separate ASR systems are trained for each of the three kinds of MFCC features. The default features computed using time-splicing followed by LDA and MLLT are employed in training and testing. Static features refer to the case of using the conventional non-adaptive approach of computing the cepstral coefficients. It is to note that the recognition performance obtained by using the static MFCC features is quite poor compared to the pitch-adaptive approaches. Further, pitch-adaptive features are effective even in the case of SGMM- and DNN-based ASR systems.

As discussed earlier, the feature normalization techniques are quite effective in the case of children's mismatched ASR task. Moreover, these techniques are found to be resulting in additive reduction in the WERs as shown in Figure 7.1. Consequently, we re-explored the effect of the fMMLR and the VTLN on the pitch-adaptive features in the context of the SGMM- and the DNN-based systems. The WERs obtained using the static features as well as the proposed pitch-adaptive MFCC features are given in Table 7.5. Also enlisted are the relative reductions in WERs due to the improved feature extraction technique. Large reductions are noted in the case of the GMM-based system for proposed pitch-adaptive features. On the other hand, for the SGM- and DNN-based systems, though the reductions are somewhat lesser, the changes are quite significant. Furthermore, the fMLLR and the VTLN lead to further reduction in the WERs in all the studied cases.

In Section 7.3, the effect of low rank feature projections on the static features was presented. We also performed a similar study on the proposed pitch-adaptive features. The WERs for this study are given in Table 7.6. In the table, static and adaptive refer to the cases of employing the conventional and the pitch-adaptive approaches for computing the base MFCC features, respectively. The 39-dimensional features obtained by the time-splicing of the base MFCC

Table 7.5: Relative improvements in recognition performances obtained through the use of the proposed pitch-adaptive MFCC features. The WERs are reported for the three acoustic modelling approaches explored along with the feature normalization using the fMLLR/VTLN.

Acoustic modelling approach	WER (in %)								
	No normalization			VTLN			fMLLR		
	Static	Prop.	PRI	Static	Prop.	PRI	Static	Prop.	PRI
GMM	62.55	50.78	19	35.06	27.62	21	43.53	35.83	18
SGMM	54.43	48.35	11	26.12	24.57	6	31.89	27.88	13
DNN	43.32	**39.32**	9	23.57	**21.43**	9	24.25	**22.33**	8

Table 7.6: Relative reductions in the WERs obtained through low-rank projection employed on the static as well as the proposed pitch-adaptive cepstral features. The WERs are being given for the three acoustic modelling approaches explored. Also given are the WERs for the cases when the fMLLR- or the VTLN-based normalization is applied along with the SW.

Acoustic modelling approach	Employed feature kind	WER (in %)								
		No normalization			VTLN			fMLLR		
		Def.	SW.	PRI	Def.	SW	PRI	Def.	SW	PRI
GMM	Static	62.55	45.59	27	35.06	26.04	26	43.53	31.96	26
	Adaptive	50.78	44.90	12	27.62	24.17	12	35.83	29.83	17
SGMM	Static	54.43	41.01	24	26.12	23.06	12	31.89	25.33	20
	Adaptive	48.35	40.65	16	25.57	23.26	9	27.88	25.75	8
DNN	Static	43.32	39.07	10	23.57	21.40	9	24.25	21.44	11
	Adaptive	39.32	36.22	8	21.43	20.07	6	22.33	20.31	9

features followed by LDA and MLLT are being referred to as the default in Table 7.6. From the reported performances, it can be noted that the low-rank projection approach (SW) is effective even in the case of the adaptive features. The relative reductions in the WERs due to the feature projection turn out to be smaller in the case of adaptive features in comparison to the static case. Similar trends are noted for all the three acoustic modelling techniques. This hints that there has been a saturation in performance to a certain extent.

To further build our confidence in the trends noted so far, we repeated the experimental evaluations for the mismatched ASR on wideband speech data, i.e., sampled at 16 kHz rate. The WERs for this study are given in Table 7.7. The obtained WERs are found to be consistent

Table 7.7: Relative improvements in the recognition performances obtained through the proposed adaptive-liftering-based MFCC features in combination with existing speaker normalization techniques. These evaluations are done separately in the context of the three different acoustic models trained on the speech data sampled at 16 kHz rate.

Acoustic modelling approach	WER (in %)					
	VTLN			fMLLR		
	Static	Adaptive	PRI	Static	Adaptive	PRI
GMM	28.57	24.15	15	35.08	29.54	16
SGMM	25.80	23.10	10	24.72	22.21	10
DNN	21.76	**19.83**	9	20.38	**18.14**	10

with those obtained in the case of narrowband speech data.

7.4.1 Adaptive-liftering-based Mel-filterbank features

Some of the recent works have shown that the use of log-compressed energies at the output of Mel-filterbank as features in the case of DNNs outperform the MFCC features [6, 124, 125]. Motivated by those works, we trained a DNN-HMM system using the filterbank features as well. The number of filterbank coefficients is chosen to be 40 as suggested in [6]. The WERs for this study are given in Table 7.8. For contrast, the WERs are also given for the wideband speech case. Moreover, the WERs obtained by using PLP features are also enlisted. It is to note that the use of the filterbank features did not result in significant changes in the recognition performances.

The filterbank features are derived by following the same procedure as that for the MFCCs except that the DCT is avoided in the case of the former. Consequently, the aforementioned pitch-induced distortions affect the filterbank features as well. Motivated by this, we applied the adaptive-liftering approach to derive the filterbank features. The block diagram depicting the approach for extracting adaptive-liftering-based filterbank features is shown in Figure 7.4.

A study on the effect of spectral smoothening through pitch-adaptive liftering in the case of MFCC features is already reported in Section 6.3. It is expected that the variances of the filterbank features would be affected similarly. But, it would be better to validate the same. For this purpose, a study was performed on the same vowel data that has been described in Section 4.1. Figure 7.5 shows the variances for the 23-dimensional Mel-filterbank features with and without adaptive-liftering-based spectral smoothing being applied for two broad pitch

Table 7.8: Recognition performances obtained through the default MFCC features and the proposed adaptive-liftering-based acoustic features on the DNN-HMM system along with fMLLR-based speaker normalization. The WERs are also given for the case when the log Mel-filterbank energies (Fbank) are used as features along with fMLLR in the training of the DNN-HMM system. The recognition performances are listed for the cases of speech data being sampled at 8 kHz and 16 kHz rates.

Acoustic model kind	Data used for SI system training	WER (in %)				
		Static			Adaptive	
		MFCC	Fbank	PLP	MFCC	Fbank
SGMM	Narrowband	31.89	28.56	30.26	27.88	26.44
	Wideband	23.94	24.54	24.28	21.79	21.36
DNN	Narrowband	24.25	23.90	24.00	22.33	22.18
	Wideband	20.38	19.82	20.10	18.14	18.27

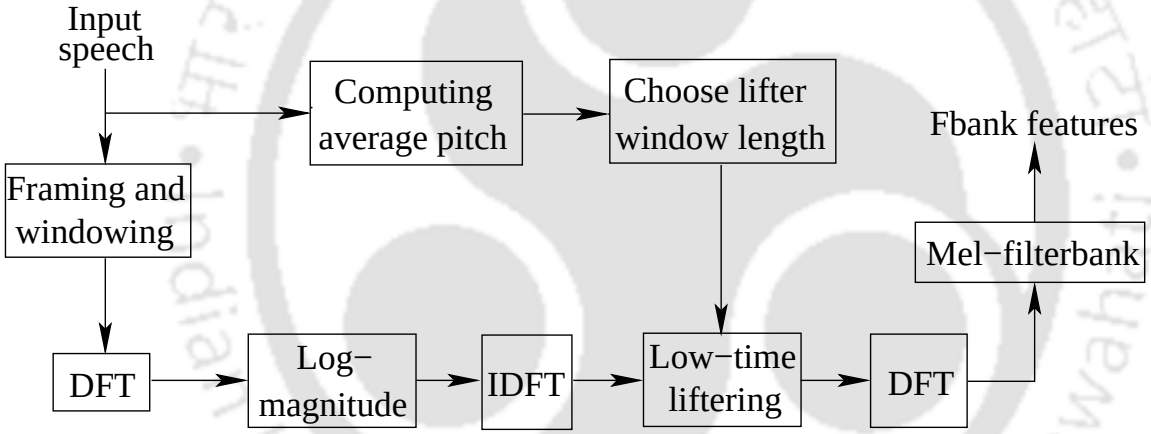

Figure 7.4: Block diagram depicting the procedure for extracting the adaptive-liftering-based filterbank (Fbank) features.

groups. The reduction in the variance mismatch between the low- and high-pitched groups, in particular for the lower index filters, is very much evident from the figure. This reduction in variance mismatch is expected to enhance the recognition performance as noted in the case MFCC features. The WERs obtained by the use of adaptive-liftering based pitch-robust features are also given in Table 7.8. It is to note that, for the children's mismatched ASR, the use of the proposed pitch-adaptive features is found to be significantly better than using the static filterbank features.

The filterbank features happen to be highly correlated. Consequently, the development of the GMM-based system using the filterbank features will require the use of full covariance

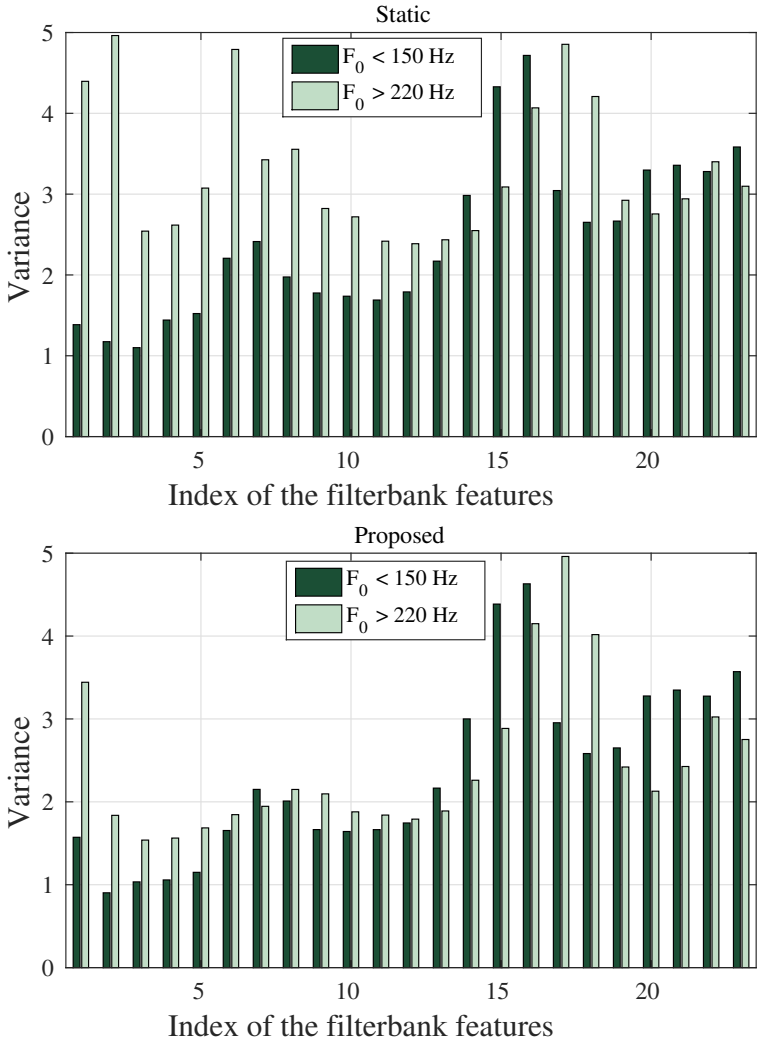

Figure 7.5: Variance of the log-compressed Mel-scaled filterbank features for the vowel /IY/ for two broad pitch (F_0) ranges. Figure also shows the reduction in the variance mismatch as a result of the pitch harmonic smoothening achieved by the proposed approach.

matrices in modelling. This would lead to substantial increase in the complexity. In the case of the SGMM-based acoustic models, on the other hand, the HMM states are modeled using shared full covariance matrices. Therefore, we explored the role filterbank features in the context of the SGMM-based mismatched ASR task. The static as well as the adaptive-liftering-based filterbank features are employed for this study and the corresponding WERs are given Table 7.8. Its quite evident from the enlisted WERs that the use of proposed filterbank is effective even in the case of SGMM-based ASR system.

7.5 Summary

In this chapter, we have explored the recently developed SGMM- and DNN-based acoustic modelling techniques in the context of children's mismatched ASR task. Severe degradations in the recognition performances are noted even with the use of these acoustic modelling techniques. Consequently, we explored the low-rank feature and the pitch-adaptive MFCC features for enhancing the mismatched system performance. Both the techniques are found to be highly effective for the mismatched ASR task. For the sake of completeness, we have included a study on the wideband band speech as well. In addition to that we explored the effectiveness of the Mel-scaled log filterbank features in the context of SGMM- and DNN-based systems. To address the issue of insufficient spectral smoothening, filterbank features are derived using the adaptive-liftering approach as well. The use of pitch-adaptive filterbank is found to be superior to that of the static case.

8

Conclusions and Future Directions

Contents

8.1 Summary and Conclusions

The work presented in this thesis explores some of the challenges in recognizing children's speech on ASR systems trained using data from adult speakers. A severely degraded recognition performance is generally observed in such tasks due to the gross mismatch in the acoustic attributes between the two groups of speakers. Improving the recognition of children's speech in acoustically mismatched conditions is quite desirable since there are a number of ASR applications which are accessed by both adults and children. One of the ways to achieve good ASR performance for both adults and children is to pool a large amount of data from both group of speakers in training of the system [121]. There is a scarcity of publicly available speech corpus from child speakers. Pooling a limited amount of adults' data with children's training set was not found to be very effective in the case of GMM-HMM-based ASR systems [126]. In contrast, subtle improvements were observed for the DNN-HMM-based system in that work. In our exploration, it was observed that adding a small amount of children's data to adults' training set degraded the recognition performance for the adult speakers, even though a significant improvement was noted for the children's speech. Consequently, the attempts made in this thesis explore the possibility of achieving improved recognition performance for the children from an ASR system without affecting the same for the adults.

In order to enhance the performance of a children's mismatched ASR, we have explored some of the existing adaptation/normalization techniques in the beginning of the thesis. Even though the explored approaches are found to be very effective, yet a large gap remains between the adults' matched and children's mismatched testing cases. Motivated by that, we have attempted to develop techniques that specifically target the issues in the children's mismatched ASR case. Among the various sources of mismatch identified in literature, the differences in the vocal tract dimensions and the pitch are shown to be the most dominant ones. The VTLN-based frequency warping is already known to be very effective in mitigating the ill-effects of the differences in the vocal tract dimensions. Therefore, the techniques that target the pitch variation across speakers are developed in this thesis. The proposed techniques are found to give additive gains when combined with the VTLN. Further, we have tried to reduce the latency in implementation of the developed adaptation approaches as low as possible. This makes the proposed techniques suitable for those ASR tasks that involve human-machine interactions. The developed techniques are observed to be very effective not only in the case of GMM-HMM-based ASR systems but also in the case of SGMM- and DNN-based systems.

The conclusions drawn from the analyses and the results of the experimental studies presented in the thesis are summarized in the following:

In Chapter 3, the effectiveness of existing model-interpolation-based fast adaptation techniques for improving the recognition performance of children's mismatched ASR is explored. In this regard, a novel model-space-based fast adaptation approach is also presented for adapting the mean parameters of the GMM-HMM-based SI system. The proposed fast adaptation approach exploits the sparse representation of test mean supervector over the exemplar and learned speaker dictionaries for deriving the Gaussian mean parameters of the adapted model. Furthermore, the challenges faced in linking the sparse coding process with the model interpolation have been highlighted. The developed fast adaptation technique is found to be effective not only for the children's mismatched case but also for the adults' match case. In addition to that, the proposed adaptation approach is observed to reduce the computational cost significantly in comparison to the existing techniques. For the children's mismatched case, the combination of VTLN and the proposed approach results in a relative reduction in WER by 20% over the VTLN included baseline.

As already stated, a large difference remains between the adults' matched and children's mismatched testing cases despite the use of adaptation as well as normalization techniques. From the earlier reported works on children's mismatched ASR, the observed degradation is mainly attributed to the differences in the dimensions of the vocal organs and the pitch of the two groups of speakers. The use of VTLN largely addresses the ill-effects of the differences in the vocal tract dimensions as already reported in literature. Even after the inclusion of VTLN, the observed WERs for the children's mismatched ASR are noted to be much poorer. This motivated us to explore the role of pitch-induced distortions during the extraction of front-end acoustic features and their subsequent effects on the trained acoustic models.

In Chapter 4, we analyze why the pitch-induced distortions appear during the MFCC feature extraction process. A detailed discussion from the perspective of signal processing is presented. The analysis performed on several vowels reveals that the pitch harmonics do not get sufficiently smoothed by the filterbank in the case of high-pitched child speakers. Thus leading to increase in the variances of the MFCC feature coefficients in contrast to those for the low-pitched adult speakers. This accounts for the poor likelihoods obtained for the children's test speech with respect to the acoustic models trained on adults' speech. Consequently, the recognition performances for children's mismatched ASR turn out to be highly degraded as noted in Chapter 3.

Following the analysis, in order to address the mismatch in the variances of the acoustic features due to differences in the pitch, a structured low-rank feature-projection-based fast adaptation approach is presented in Chapter 5. The developed technique intends to map the training as well as the test data to a lower dimensional subspace such that the variance mismatch is effectively reduced. From the analysis as well as the experimental evaluations presented in this chapter, it may be concluded that any dimensionality reduction technique that is essentially based on the variance of the data, can be used for deriving the low-rank feature projections. Two different dimensionality reduction techniques, viz. HLDA and PCA have been explored in this thesis to substantiate this claim. Both HLDA and PCA are observed to highly effective in mitigating the ill-effects of pitch-induced distortions. The same is validated experimentally in this chapter and is noted to significantly reduce the WERs. Further, we have also shown that employing a constrained block-diagonal structure is more effective than learning full HLDA/PCA transforms. The introduced constrained block-diagonal structure not only improves the recognition performance but also reduces the computational cost involved in deriving the low-rank feature projections. In addition to that, the low-rank feature projection is effectively combined with various model-space adaptation as well as feature-space normalization techniques to achieve added gains in the recognition performance. A relative reduction of 35% in WER over the baseline is obtained with the low-rank feature projection. An added relative reduction of 14% is achieved when the proposed fast-adaptation approach is included. Projecting the trained model parameters as well as the test data to a lower dimensional subspace is found to be effective in the case of PLPCC features also. A relative reduction of 35% in WER over the baseline is observed in the case of PLPCC features as well.

The widely used MFCC feature extraction process involves static signal processing which leads to the appearance of pitch-induced distortions as discussed in Chapter 4. Consequently, we have explored the role of pitch-adaptive signal processing in the extraction of MFCC features in Chapter 6. In this regard, a simple technique based on pitch-adaptive liftering is proposed. The use of pitch-adaptive MFCCs is noted to be superior to that of the static MFCCs in terms of recognition performance. Further, we have also compared the pitch-adaptive features with other existing acoustic feature extraction process (like PLPCC) to gauge their relative effective in the case of mismatched ASR task. We have also explored the effectiveness of combining the existing feature normalization techniques and the low-rank feature projection with the pitch-adaptive MFCCs. When combined with fMLLR/VTLN-based normalization, a relative reduction of 18%/21% in WER over the static MFCC case is noted. A further relative reduction of 12% is

obtained when pitch-adaptive features are projected to a lower dimensional subspace. At the same time, the use of proposed acoustic features does not degrade the recognition performance in the case of adults' matched ASR task.

Finally, the effectiveness of low-rank projection of proposed pitch-adaptive MFCCs is also explored on ASR systems employing the SGMM- and the DNN-based acoustic modeling in Chapter 7. Like the GMM-HMM case, ASR systems are developed for the adult speakers employing SGMM- and DNN-based acoustic modelling, respectively. This is followed by incorporating the earlier developed techniques for improving the recognition performance for the child speakers. Both the explored approaches, viz. the low-rank feature projection and the pitch-adaptive MFCC features are found to be effective even in the case of SGMM/DNN-based ASR systems. The combination of the two yields a relative reduction of 9%/6% for the SGMM/DNN-based system over their respective baselines. In addition to that, Mel-scaled filterbank feature derived using the proposed pitch-adaptive spectral smoothening approach is also explored in the context of SGMM- and DNN-based systems. The pitch-adaptive filterbank features are noted to be superior to the static ones.

It is to note that, even though the use of DNN-based acoustic modelling has been extensively explored for adults' matched ASR tasks, not much work has been done in the mismatched case. As mentioned earlier, only a few works on children's ASR employing DNN-based acoustic modelling have been reported [119–121]. From the studies reported in this chapter, the following conclusions can be drawn:

- Though SGMM/DNN-based systems are superior to the GMM ones, the ill-effects of pitch-dependent distortions are still evident in the case of mismatched ASR;
- The proposed approaches are found to be effective even in the case of SGMM- and DNN-based children's mismatched ASR systems;
- Unlike the adults' matched task, the VTLN is as effective as the fMLLR;
- The VTLN can be cascaded with fMLLR for additive gain in the recognition performance;
- The VTLN is found to be effective for both shallow and deep networks unlike that noted for the adults' matched tasks.

8.2 Scope of the Future Work

Some recent works have explored deep convolutional neural networks (CNN) [127] as well as deep recurrent neural networks (RNN) [128] for speech recognition. In future we wish to extend the techniques presented in this thesis on to CNN/RNN-based systems. Since model-space adaptation approaches were observed to be effective in combination with the proposed approaches, we wish to implement the same in the DNN/CNN domain as well. Adaptation of DNN/CNN-based ASR system has also been explored in the last few years [129–132]. Among those, the cluster adaptive training for DNN [133] is a fast adaptation approach. Following that study, the fast adaptation techniques reported in this thesis could be extended to DNN-based systems.

A number of works using bottleneck (BN) features has been reported for speech recognition [134, 135]. In earlier works the BN features for speech were created from a multi-layer perceptron (MLP) trained to predict the phonemes or the phoneme states. With the recent advancements, the BN features are being generated from a DNN-based system for classifying senones [136, 137]. Combining BN features with other complementary features like MFCCs/PLPs to train a GMM-HMM system is reported to outperform the system trained using MFCCs/PLPs [138]. Generally the combination of the BN and the MFCC features are processed via PCA/HLDA for dimensionality reduction and decorrelation before modelling. In future, we wish to extract BN and tandem features using the proposed pitch-adaptive cepstral features as the input to the DNN.

Another challenging problem in speech processing is the automatic recognition of speech under stress. A few studies on DNN-based ASR for stressed speech are reported [139–141]. The effect of stress is more pronounced in the low frequency region of the stressed speech in comparison to the high frequency region as observed in some of the earlier reported works [142, 143]. This leads to a similar nature mismatch in the variance of the acoustic features between the neutral and stressed speech conditions. It is hypothesized that the low-rank feature projections developed for addressing pitch mismatch should turn out to be effective in the case of stressed speech recognition as well. In an initial exploration, the low-rank projections are noted to achieve 10-30% reduction in the WER for speech belonging to varying stress classes, thus supporting our hypothesis.

A

Front-end Acoustic Features

Contents

A.1 Mel-frequency Cepstral Coefficients

One of the most commonly used acoustic features in speech recognition is the Mel-frequency cepstral coefficients (MFCC) [3]. The typical MFCC feature extraction process consists of the following steps: windowing of the speech signal, applying the discrete Fourier transform (DFT) on the windowed signal to derive the the magnitude spectrum for each frame, taking the log of the magnitude and then warping the frequencies on to the Mel-scale and finally applying the inverse discrete cosine transform (DCT) followed by low-time liftering. In the following we discuss each of the steps involved in the extraction of the MFCC features in brief.

(i) **Pre-emphasis:** Pre-emphasis is the filtering process that emphasizes the higher frequency components of the speech signal. This is done in order to balance the spectrum of voiced sound units having a steep roll-off in the high frequency region. In general, the speech signal is observed to have an overall spectral slope of -6 dB per octave. This net roll-of is due to the combined effect of the glottal pulse roll-off (-12 dB/octave [144]) and the lip radiation ($+6$ dB/octave). The transfer function for the most commonly used pre-emphasis filter is as following:

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

where the slope of the filter is controlled by using the parameter b whose value usually lies between 0.9 to 1.0 [144].

(ii) **Frame blocking and windowing:** It is well known that the speech signal is a slowly varying or a quasi-stationary signal with respect to time. Therefore, the analysis of speech is always carried out on short segments across which the speech signal is assumed to be stationary. Typically, an analysis window of size 20-25 ms is considered for the short-time spectral analysis. The duration of the window is still long enough to contain at least one cycle of the lowest frequency component of interest. Usually, overlapping frames with an overlap of 10 ms are considered. This is done in order to enable the tracking of the temporal characteristics of the individual speech sounds. In order to force the values near the edges of the frames to be zero, each frame is multiplied by a tapered window, viz. the Hanning or the Hamming window [144]. This enhances the harmonics, smooths the edges and reduces the discontinuities at the edges while taking the DFT on the signal.

(iii) **Computing the DFT spectrum:** After windowing, each speech frame is converted into its frequency domain representation by applying the DFT. For any signal $x(n)$, the DFT $X(k)$ is given as follows [145]:

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

where N is the number of points used to compute the DFT. The phase information is discarded from the resulting short-term spectrum of a speech frame. The phase of the spectrum is discarded because it does not carry any useful information from the perspective of hearing.

(iv) **Computing the Mel-spectrum:** Next, the Fourier transformed signal is passed through a set of band-pass filters known as the Mel-filterbank to compute the Mel-spectrum. A Mel is a unit of measurement based on the human ear's perceived frequency. It does not correspond linearly to the physical frequency of the tone. The Mel-scale has approximately a linear frequency spacing below 1 kHz and a logarithmic spacing above 1 kHz [146]. The approximation of the Mel from the physical frequency can be expressed as:

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

where f denotes the physical frequency in Hz and f_{Mel} denotes the perceived frequency [100].

The employed filterbanks can be implemented either in the time domain or in the frequency domain. In general, the filterbanks are implemented in the frequency domain for the computation of the MFCC features. The center frequencies of the filters are normally evenly spaced on the frequency axis. In order to mimic the human ears' perception, the warped axis is implemented according to the non-linear function given in (A.3). The bandwidths of the filters are decided on the basis of the critical bandwidth phenomena noted in the psychoacoustic studies for the human auditory perception. Therefore, each pair of consecutive filters have an overlap of 50% [147]. The most commonly used filter shaper is triangular in shape. In some cases the Hanning filter is also found to be employed [144].

The Mel-spectrum of the magnitude spectrum $X(k)$ is computed by multiplying the

magnitude spectrum by each of the of the triangular Mel-weighted filters as follows:

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

where M is the total number of triangular Mel-weighting filters [148]. $H_m(k)$ is the weight given to the k^{th} energy spectrum bin contributing to the m^{th} output band and is expressed as:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{2(k-f(m-1))}{f(m)-f(m-1)}, & f(m-1) \leq k \leq f(m) \\ \frac{2(f(m+1)-k)}{f(m+1)-f(m)}, & f(m) < k \leq f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{A.5}$$

with m ranging from 0 to $M-1$.

(v) **Discrete cosine transform (DCT):** Since the vocal tract is smooth, the energy levels in the adjacent bands tend to be correlated. The DCT when applied to the logarithm of the transformed Mel-frequency coefficients produces a set of de-correlated cepstral coefficients. It is to note that the DCT gathers most of the information in the signal to its lower order coefficients. Consequently, by discarding the higher order coefficients, significant reduction in computational cost and robustness of systems can be achieved [144]. This step of discarding the higher-order coefficients is referred to as the low-time liftering. Typically, for the task of speech recognition, the first 13 coefficients are chosen and are referred to as the base MFFC feature. The base MFCCs are computed as per the following relation [144]:

$$c(n) = \sum_{m=0}^{M-1} \log_{10}\left(s(m)\right) \cos\left(\frac{\pi n(m-0.5)}{M}\right); \qquad n = 0, 1, 2, ..., C-1 \tag{A.6}$$

where $c(n)$ are the cepstral coefficients and C is the dimensionality of the MFCC feature vectors. The logarithm of the energy/power for the entire frame is also appended to the Mel-frequency cepstral coefficients, i.e., the zeroth order coefficient. This is done so because the different phonemes may have different energy.

(vi) **Computing the dynamic MFCC features:** The base cepstral coefficients computed following the above describe procedure are usually referred to as the static features, since

they contain information from a given frame only. The extra information about the temporal dynamics of the signal is obtained by computing the first- and the second-order temporal derivatives of the cepstral coefficients [149–151]. The first-order derivative is called the delta coefficient while the second-order derivative is termed as the delta-delta coefficient. The delta coefficients contain the information about the speech rate. The delta-delta coefficients, on the other hand, give an information similar to the acceleration of the speech. The dynamic MFCC parameters are generally computed as follows [149]:

$$\Delta c_m(n) = \frac{\sum_{i=-T}^{T} k_i c_m(n+i)}{\sum_{i=-T}^{T} |i|} \tag{A.7}$$

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 the computation. Generally T is taken to be between 2-5. The delta-delta coefficients are computed by taking the first-order derivative of the delta coefficients in the same way.

A.2 Perceptual Linear Prediction

The perceptual linear prediction (PLP) is another commonly used acoustic feature in the ASR systems [4]. The PLP approach employs several concepts from the psychophysics of hearing for frequency weighting to determine an estimate of the auditory spectrum. An autoregressive low-order all pole model is then employed to approximate the auditory spectrum. As a result, the PLP spectrum does not reflect the speaker-dependent details in the spectrum of the speech, merging the higher resonance spectral peaks. The basic steps involved in a typical perceptual linear prediction cepstral coefficient (PLPCC) features computation process are described in the following.

(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, as follows:

$$W(n) = 0.54 + 0.46\, cos\left[\frac{2\pi n}{(N-1)}\right] \tag{A.8}$$

where N is the length of the Hamming window. Like the MFCC feature computation process explained earlier, the short-time spectral measurements are carried out using a analysis window of size 20-25 ms such that the frames are overlapping with their centers being only 10 ms apart. After the windowing, the discrete Fourier transform (DFT) is used to convert the windowed speech frame to its frequency domain representation as given in (A.9) [145].

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nk}{N}}; \qquad 0 \leq k \leq N-1 \tag{A.9}$$

In (A.9), N is the number of points used to compute the DFT. Next, to get the short-term power spectrum, the real and imaginary components of the short-term speech spectrum are squared and added as per the following relation:

$$P(\omega) = \mathcal{R}e[S(\omega)]^2 + \mathcal{I}m[S(\omega)]^2. \tag{A.10}$$

(ii) **Critical Band Analysis:** From the perspective of the hearing mechanism, the use of the critical bands for speech analysis is important. For a given center frequency, the critical band 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 all have the same sound pressure. Therefore, the critical bandwidth is used to represent the power of the human ears for resolving simultaneous tones. After finding the power spectrum $P(\omega)$, it is warped along its frequency axis Ω into the Bark frequency [152] as follows:

$$\Omega(\omega) = 6 \ln \left[\frac{\omega}{120\pi} + \left[\left[\frac{\omega}{120\pi} \right]^2 + 1 \right]^{0.5} \right] \tag{A.11}$$

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 masking curve $\Psi(\Omega)$ [152] using (A.12).

$$\Psi(\Omega) = \begin{cases} 0 & \Omega < -1.3, \\ 10^{2.5(\Omega+0.5)} & -1.3 \leq \Omega \leq -0.5, \\ 1 & -0.5 < \Omega < 0.5, \\ 10^{2.5-1.0(\Omega-0.5)} & 0.5 \leq \Omega \leq 2.5, \\ 0 & \Omega \geq 2.5 \end{cases} \tag{A.12}$$

The samples of the critical-band power spectrum is given by the discrete convolution of $P(\omega)$ with $\Psi(\Omega)$ as per the following:

$$\Theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i)\Psi(\Omega). \tag{A.13}$$

In comparison to the original $P(\omega)$, the convolution reduces the spectral resolution of $\Theta(\Omega)$. Down-sampling of the Bark scale spectrum $\Theta(\Omega)$ is then performed by re-sampling every one Bark. Typically, 18 spectral samples of are used to cover the 0-16.9 Bark (0-5 kHz) analysis bandwidth in 0.994 Bark steps.

(iii) **Equal Loudness Pre-emphasis:** Pre-emphasis of the sampled $\Theta[\Omega(\omega)]$ is performed using the simulated equal loudness curve given by the following equation:

$$\Xi[\Omega(\omega)] = E(\omega)\Theta[\Omega(\omega)] \tag{A.14}$$

where the function $E(\omega)$ is given by (A.15)

$$E(\omega) = \frac{\omega^4(\omega^2 + 56.8 \times 10^6)}{(\omega^2 + 6.3 \times 10^6) \times (\omega^2 + 0.38 \times 10^9)}. \tag{A.15}$$

The function $E(\Omega)$ is an approximation to the nonequal sensitivity of the human hearing at different frequencies (adopted from [153]). This equation is used to simulate the resolution power of hearing. Furthermore, the above equation is the transfer function of a filter having asymptotes of 12 dB/octave between 0 and 400 Hz, 0 dB/octave between 400 and 1200 Hz, 6 dB/octave between 1200 and 3100 Hz, and 0 dB/octave between 3100 Hz and the Nyquist frequency. This approximation is well up to 5000 Hz. For applications requiring Nyquist frequency greater than 5000 Hz, (A.15) is modified as follows:

$$E(\omega) = \frac{\omega^4(\omega^2 + 56.8 \times 10^6)}{(\omega^2 + 6.3 \times 10^6) \times (\omega^2 + 0.38 \times 10^9) \times (\omega^6 + 9.58 \times 10^{26})}. \tag{A.16}$$

Finally, since the samples at 0 Bark and the Nyquist frequency are not well defined, those are made equal to the values of their nearest neighbors. Consequently, the function $\Xi[\Omega(\omega)]$ begins and ends with two equal-valued samples.

(iv) **Intensity Loudness Power-Law:** After applying the PLP filtebank, cubic root amplitude compression is performed to approximate the power law of hearing as suggested in [154]. This compression is given in the following:

$$\Phi(\Omega) = \Xi(\Omega)^{0.33} \tag{A.17}$$

The above given operation is performed in order to simulate the nonlinear relation between the intensity of the sound and its perceived loudness. Furthermore, it is also effective in reducing the spectral amplitude variation of the critical band spectrum when applied together with the psychophysical equal-loudness pre-emphasis filter. Consequently, these two approaches allow a lower-order all-pole modeling, discussed next, with a reduced computational cost.

(v) **Autoregressive Modeling:** Finally, $\Phi(\Omega)$ is approximated by the spectrum of an all-pole model. This is done using the autocorrelation method of an all-pole spectral modeling. The autocorrelation function dual of $\Phi(\Omega)$ is obtained by applying the inverse DFT (IDFT). Typically, a 34-point IDFT is used. Furthermore, the IDFT is a better choice

than the inverse FFT as only a few autocorrelation values are needed. Next, the Yule-Walker equations for the autoregressive coefficients of the M^{th} order all-pole model are solved using the first $M+1$ autocorrelation values.

(vi) **PLP Cepstral Coefficients:** The PLP coefficients obtained in the last step above are then employed to compute PLP cepstral coefficient (PLPCC) using the recursion formula suggested in [155]. Those recursion formulae are given as follows:

$$c_1 = a_1 \tag{A.18}$$

$$c_n = \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) a_k c_{n-k} + a_n \quad 1 < n < M \tag{A.19}$$

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 using the steps defined in [149–151].

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

B

Acoustic Modelling Approaches

Contents

B.1 GMM-HMM-based Acoustic Modelling

In practical large vocabulary automatic speech recognition (ASR) systems, one of the techniques to learn the acoustic models is based on the Hidden Markov model (HMM)[1]. The observation probabilities of the HMM states are, in turn, modelled using multivariate Gaussians mixtures (GMM) [2]. In this approach, the acoustic model consists of an HMM for each of the basic modelling unit, e.g., a phone. The HMM of a word in the vocabulary is then derived by concatenating the corresponding phone-specific HMMs. The HMMs of the words can be further concatenated to construct the HMM of a sentence/phrase.

Given the an observation vector $\mathbf{o}_l$ for the frame l, the set of model parameters that defines a continuous density HMM Λ, with discrete sequence of S states that are modeled using Gaussian density functions having M mixtures, comprises of:

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

(ii) $\mathbf{B} = \{b_j(\mathbf{o}_l);\ \ 1 \leq j \leq S\}$, the observation probabilities of a state

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

In the context of the ASR systems, a left-to-right HMM topology is employed. Two non-emitting sates are included as the entry and the exit states along with the emitting states. The non-emitting states do not generate any observation unlike the emitting states. Consequently, their respective initial probabilities are $\pi_1 = 0$ and $\pi_i = 0$ for $i \neq 1$.

For a GMM-HMM-based isolated unit ASR system, the algorithms used for training (learning parameters of the model from observations) and testing (decoding a test utterance to hypothesize a word) are described in the following subsections.

B.1.1 Learning the GMM parameters

The set of model parameters of the HMMs viz., the Gaussian means, the variances of the Gaussians, the Gaussian mixture-weights and the state transition probabilities are learned in a data-driven manner using the Baum-Welch re-estimation algorithm [156]. A variation of the expectation-maximization (EM) algorithm [157] is used to learn the HMMs parameters employing maximum-likelihood training. Given an initial model Λ and the observation sequence $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T\}$, each iteration increases the likelihood of the training data i.e., $P(\mathbf{O}|\Lambda)$ till it converges to a local maximum. The function $P(\mathbf{O}|\Lambda)$ is called the likelihood function.

[1]This discussion closely follows the detail description given in [149]

At the beginning of the training process, a rough guess of the parameter values $\mathbf{A}$, $\mathbf{B}$ and π of an initial model Λ is made. This can be done using either a flat start training or the segmental K-means algorithm [158]. 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 $l = 1$.

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

where $a_{ij} = P(i_{l+1} = j | i_l = i)$, is the probability of being in state j at time $l+1$ given that the system was in state i at time l. The a_{ij}'s are assumed to be independent of time.

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

where $b_j(\mathbf{o}_l) = \sum_{m=1}^{M} c_{jm} \mathcal{N}(\mathbf{o}_l; \boldsymbol{\phi}_{jm}, \mathbf{C}_{jm})$, is the probability of observing the symbol $\mathbf{o}_l$ given the state j. The multivariate Gaussian density function for a D dimensional observation feature vectors is given as:

$$\mathcal{N}(\mathbf{o}_l; \boldsymbol{\phi}_{jm}, \mathbf{C}_{jm}) = \frac{1}{\sqrt{(2\pi)^D |\mathbf{C}_{jm}|}} \exp\left\{-\frac{1}{2}(\mathbf{o}_l - \boldsymbol{\phi}_{jm})^T \mathbf{C}_{jm}^{-1} (\mathbf{o}_l - \boldsymbol{\phi}_{jm})\right\} \tag{B.4}$$

where, c_{jm}, $\boldsymbol{\phi}_{jm}$ and $\mathbf{C}_{jm}$ are the mixture-weight, the mean and the covariance matrix for the m^{th} Gaussian component in the j^{th} state, respectively.

In flat start training, on the other hand, all models are initialized to be identical. The Gaussian means and the covariances of each state are equal to the global mean and variance of the 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 $\mathbf{O} = \mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_L$ and the model $\Lambda = (\mathbf{A}, \mathbf{B}, \pi)$, subject to the stochastic constrains:

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

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

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

$$\int_{-\infty}^{\infty} b_j(\mathbf{o}_l)\mathrm{d}\mathbf{o}_2 = 1; \quad \text{for } 1 \leq j \leq S, \tag{B.8}$$

the Baum-Welch re-estimation formulae are used to get more accurate parameter values as follows:

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

$$\hat{a}_{ij} = \frac{\sum_{l=1}^{L-1} \xi_l(i,j)}{\sum_{l=1}^{L-1} \gamma_l(i)} = \frac{\sum_{l=1}^{L-1} \alpha_l(i) a_{ij} b_j(\mathbf{o}_l)\beta_l(j)}{\sum_{l=1}^{L-1} \alpha_l(i)\beta_l(i)} \tag{B.10}$$

$$\hat{c}_{jm} = \frac{\sum_{l=1}^{L-1} \gamma_l(j,m)}{\sum_{l=1}^{L-1}\sum_{m=1}^{M} \gamma_l(j,m)} \tag{B.11}$$

$$\hat{\boldsymbol{\phi}}_{jm} = \frac{\sum_{l=1}^{L-1} \gamma_l(j,m).\mathbf{o}_l}{\sum_{l=1}^{L-1} \gamma_l(j,m)} \tag{B.12}$$

$$\hat{\mathbf{C}}_{jm} = \frac{\sum_{l=1}^{L-1} \gamma_l(j,m).(\mathbf{o}_l - \boldsymbol{\phi}_{jm})(\mathbf{o}_l - \boldsymbol{\phi}_{jm})^T}{\sum_{l=1}^{L-1} \gamma_l(j,m)} \tag{B.13}$$

where

$$\gamma_l(i) = P(i_l = i|\mathbf{O}, \Lambda) \tag{B.14}$$

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

$$\gamma_l(j,m) = \left[\frac{\alpha_l(j)\beta_l(j)}{\sum_{j=1}^{S} \alpha_l(j)\beta_l(j)}\right]\left[\frac{c_{jm}\mathcal{N}(\mathbf{o}_l; \boldsymbol{\phi}_{jm}, \mathbf{C}_{jm})}{\sum_{m=1}^{M} c_{jm}\mathcal{N}(\mathbf{o}_l; \boldsymbol{\phi}_{jm}, \mathbf{C}_{jm})}\right] \tag{B.15}$$

denotes the probability of being in state j at time l with the observation $\mathbf{o}_l$ being generated by the m^{th} mixture component and

$$\xi_l(i,j) = P((i_l = i, i_{l+1} = j|\mathbf{O}, \Lambda) \tag{B.16}$$

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

$$\sum_{l=1}^{L-1} \gamma_l(i) = \text{Expected number of transitions from state i} \tag{B.17}$$

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

Therefore, the re-estimation formula for π_i is the probability of being in state i at time l. On the other hand, 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 Search and Decoding

Given observation sequence $\mathbf{O}$, the GMM-HMM system determines the most likely model Λ using the Baye's rule as follows:

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

where, $P(\Lambda)$ is 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 same for all observations. In other words, one is required to determine the maximum value of the product of $P(\Lambda)$ and $P(\mathbf{O}|\Lambda)$. This, in turn, 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 $P(\Lambda)$. In decoding, therefore, the problem is to find a state sequence $I = i_1, i_2, \ldots, i_T$ such that the joint probability of the observation sequence $\mathbf{O} = \mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T$ and the state sequence given the model $\Lambda = (\mathbf{A}, \mathbf{B}, \pi)$ is maximized. A dynamic programming algorithm known as the Viterbi algorithm [156] is used to identify the underlying hidden state sequence I that maximizes $P(\mathbf{O}, I|\Lambda)$. The Viterbi is an inductive algorithm in which at each instant the state sequence giving the maximum probability for each of the N states is kept as the intermediate state for the desired observation sequence $\mathbf{O} = \mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T$. In this way, the best path for each of the N states is determined as the last state for the desired observation sequence. Out of those the one which has the highest probability is selected.

The step-by-step implementation of the Viterbi algorithm given in [149] is reproduced in the following:

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

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

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

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

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

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

(iii) **Termination:**

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

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

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

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

Hence, the required state-optimized probability is given by $\exp(-P^*)$ while $Q^* = q_1^*, q_2^*, \ldots, q_T^*$ is the optimal state sequence.

B.2 SGMM-HMM-based Acoustic Modelling

In this section, we provide a brief discussion on the parameterization of the subspace GMM (SGMM) for ASR. This discussion closely follows the works reported in [5, 159, 160]. In the GMM-HMM-based acoustic modelling approach, a dedicated mixture of multivariate Gaussians is used to model each state level. Further, no parameters are shared between states. An alternative acoustic modelling approach reported lately in literature is the SGMM. The SGMM has similarities with the techniques like the EV- and the CAT-based speaker adaptation approaches as well as joint factor analysis (JFA) used in speaker verification [161]. In the SGMM, the model parameters represent a globally shared subspace. A set of low dimensional state-specific vectors, referred to as the *state projection vectors* $\{\mathbf{v}_j\}$, are trained from the data to capture the principal directions of the phonetic variability. The state projection vectors, in turn, determine the means and the mixture-weights of the Gaussians in the model. Moreover, the covariances in the SGMM are shared among all the HMM states and are represented by full, instead of diagonal, covariance matrices. A large portion of the SGMM parameters are dedicated to shared full covariance Gaussian subspace parameters. A relatively smaller number of parameters are used for state projection vectors. As a result of this, the SGMM facilitates acoustic modelling with a smaller amount of training data. This is thought to result partly from the subspace constraints provided by the model structure and partly due to the overall reduction in the total number of model parameters.

In the case of an SGMM-based system having J states, the observation density function of a feature vector $\mathbf{o}_l \in \mathbb{R}^D$ in state j is given by

$$b_j(\mathbf{o}_l \,|\, j) = \sum_{i=1}^{I} w_{ji}\, \mathcal{N}(\mathbf{o}_l \,;\, \boldsymbol{\phi}_{ji},\, \mathbf{C}_i) \tag{B.27}$$

where I is the number of the full covariance Gaussians that are shared between the J states. The i^{th} mean vector $\boldsymbol{\phi}_{ji}$ corresponding to the j^{th} state is derived by a projection into the i^{th} subspace as given by

$$\boldsymbol{\phi}_{ji} = \mathbf{M}_i \mathbf{v}_j \tag{B.28}$$

where $\mathbf{M}_i$ is the linear subspace projection matrix of dimensions $D \times S$. In (B.28), the state projection vector for j^{th} state is denoted by $\mathbf{v}_j \in \mathbb{R}^S$, with the subspace dimension generally being around the same to the feature dimension, i.e., $S \simeq D$. The dimension of S happens to be much smaller than the number of parameters per state. Consequently, the models span a subspace of the total parameter space and hence the name. The state specific weights in (B.27)

are derived from the state projection vector $\mathbf{v}_j$ using a log-linear model as follows:

$$w_{ji} = \frac{\exp(\mathbf{w}_i^T \mathbf{v}_j)}{\sum_{k=1}^{I} \exp(\mathbf{w}_k^T \mathbf{v}_j)} \tag{B.29}$$

where $\mathbf{w}_i$ denotes the shared subspace weight projection vectors. Additional flexibility in the SGMM parameterization is provided by introducing the notion of a substate within a state. With this modification, the distribution for the observation $\mathbf{o}_l$ in state j is given as a weighted combination of densities:

$$b_j(\mathbf{o}_l \,|\, j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \, \mathcal{N}(\mathbf{o}_l \,;\, \boldsymbol{\phi}_{jmi},\, \mathbf{C}_i) \tag{B.30}$$

where c_{jm} is the relative weight of substate m within the j^{th} state. The substate projection vectors $\mathbf{v}_{jm}$ are now used to obtain the means and the mixture-weights as follows:

$$\boldsymbol{\phi}_{jmi} = \mathbf{M}_i \mathbf{v}_{jm} \tag{B.31}$$

$$w_{jmi} = \frac{\exp(\mathbf{w}_i^T \mathbf{v}_{jm})}{\sum_{k=1}^{I} \exp(\mathbf{w}_k^T \mathbf{v}_{jm})}. \tag{B.32}$$

Multiple substate per state SGMM models are realized in multiple iterations by splitting $\mathbf{v}_{jm}$ vectors with the highest occupancy counts. Introducing the notion of the substate increases the number of parameter being estimated. Depending upon the amount of available training data, the value of m can be varied to get an optimal set of model parameters.

For model training, a single GMM called the universal background model (UBM) is learned on all speech classes pooled together. The subspace parameters $\mathbf{M}_i$, $\mathbf{v}_{jm}$ and $\mathbf{C}_i$ are initialized in such a way that the means and the variances in each state on the first iteration are the same as the UBM. The usual expectation-maximization (EM) algorithm employing a maximum likelihood criterion is used to update the parameters of the SGMM in successive iterations.

B.3 DNN-HMM-based Acoustic Modelling

The GMM-based acoustic modelling is observed to be inefficient in modelling the data that lie on or near a non-linear manifold in the data space. This is one of the major drawbacks of the GMM-HMM-based approach as suggested in [6]. On the other hand, the artificial neural network (ANN) [162] is reported to have the potential to learn these models of data that lie on or near the non-linear manifold. The lack of hardware and adequate algorithms had impaired the initial works done on the application of the ANN-based ASR system. Training an ANN with many hidden layers by the backpropagation algorithm [162] using a large amount of data was found to be infeasible. Most of the works were therefore constrained to make use of a single hidden layer which did not result in improvements significant enough to seriously challenge the GMM-HMM paradigm. Both the aforementioned limitations have been overcome with the developments made in the past few years. Efficient techniques are now available to train neural nets with a large number of hidden layers. Deep neural networks (DNN) containing many layers of non-linear hidden units and a very large output layer are now being used for modelling the acoustic variations in the ASR systems [163]. In this case, the DNN is trained to model the posterior probabilities of the senones (context-dependent tied state) which are then used in an HMM-based classifier. Such an approach is reported to outperform the ASR systems employing GMM-based acoustic modelling.

Deep neural networks are created by stacking layers of restricted Boltzmann machine (RBM) [6]. An RBM is an undirected generative model. An undirected model uses a single set of parameters ($\mathbf{W}$) to define the joint probability of a vector of observable variables ($\mathbf{v}$) and a vector of values for the latent/hidden variables ($\mathbf{h}$) via an energy function E given as:

$$p(\mathbf{v}, \mathbf{h}; \mathbf{W}) = \frac{1}{Z} e^{-E(\mathbf{v}, \mathbf{h}; \mathbf{W})} \tag{B.33}$$

where, Z is the partition function given by

$$Z = \sum_{\mathbf{v}', \mathbf{h}'} e^{-E(\mathbf{v}', \mathbf{h}'; \mathbf{W})} \tag{B.34}$$

where, $\mathbf{v}', \mathbf{h}'$ are dummy variables used for summing over the ranges of $\mathbf{v}$ and $\mathbf{h}$, respectively. In undirected models, inference (the state of the hidden unit) is easy to determine provided the hidden variables do not have interconnections among them. This property of such a restricted class of undirected models makes it ideal for pre-training a deep neural network since the output of each layer is easily inferred. An efficient unsupervised algorithm is described in [164]

for learning the connection weights in a deep belief network (DBN) that is equivalent to training each adjacent pair of layers of an RBM. There is also a fast, approximate, bottom-up inference algorithm to infer the states of all hidden units conditioned on a data vector [165].

After training an RBM on the data, the output of the hidden units can be used as the input data for training another RBM. This helps learn to model the significant dependencies between the hidden units of the first RBM. This can be repeated many times to produce many layers of non-linear feature detectors that represent more complex statistical structure in the data. For each data vector $\mathbf{v}$, the vector of hidden unit activation probabilities $\mathbf{h}$ is computed. These hidden activation probabilities are then used as the training data for a new RBM. Thus, each set of RBM weights can be used to extract features from the output of the previous layer. The initial values for all the weights of the hidden layers of the neural nets can thus be generated using the RBM training (the number of hidden layers being equal to the number of RMBs trained). This is called pre-training of a DBN. A randomly initialized softmax output layer is then added and all the weights in the network are discriminatively fine tuned using the backpropagation to create a DNN.

In the case of the ASR systems, the softmax output layer has as many node as the number of classes, i.e., the number of senones in the case of DNNs. For the multiclass classification problems like speech recognition, the output unit j converts the total input x_j into a class probability p_j by using the softmax nonlinearity given by

$$p(x_j) = \frac{\exp(x_j)}{\sum_k \exp(x_k)} \tag{B.35}$$

where k is an index over all the classes. The DNN's can be trained by backpropagating derivatives of the cost function i.e., the cross-entropy between the target probabilities and the output of the softmax function. The final fine tuning phase is supervised and hence the labeled data is required. The DNN-HMM architecture is used to model the senones directly as the output of the DNN. The input speech is fed through the input layer $\mathbf{v}$ while the output of the DNN is the scaled likelihood of the senones $p(x_j)$. This senone likelihood is used with HMM which models the sequential property of the speech.

Some of the reported works have shown that DBN pre-training is not mandatory when the available data for training is sufficiently large [6] or when a very large number of hidden layers is employed (more than 7) [123]. In such cases, one can use the DNN training approach reported in [166]. This approach is very similar to the greedy layer-wise supervised training [167] or the "layer-wise backpropagation" reported in [168]. In this approach, the network is initialized

randomly with one hidden layer which is then trained for a short time, i.e., typically less than an epoch or in other words less than one full-pass through the data. Next, the layer of weights that go to the softmax layer is removed, a new hidden layer and two sets of randomly initialized weights are added and training is done again. This is repeated until the desired number of layers are trained. The Kaldi speech recognition toolkit supports both the kinds of DNN-HMM training.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

C

Deriving Minimum-Phase System Response

For computing the minimum-phase frequency response from the magnitude, one of the ways is the cepstral method as given in [169]. The involved steps are summarized as follows:

(i) Given the Mel-spectra, obtain the the real cepstrum, $c(n)$.

(ii) The noncausal portion of $c(n)$ is folded onto its causal portion.

(iii) Perform a forward FFT, followed by exponentiation to obtain the minimum phase frequency response Smp.

These steps are executed by the MATLAB code give below:

```
load('mel_spectra.mat');

NZ = 1;       % number of ZEROS in the filter to be designed
NP = 8;       % number of POLES in the filter to be designed
NG = 10;      % number of gain measurements
fmin = 100;  % lowest measurement frequency (Hz)
fmax = 3000; % highest measurement frequency (Hz)
fs = 8000;  % discrete-time sampling rate
Nfft = 200;  % FFT size to use
df = (fmax/fmin)^(1/(NG-1)); % uniform log-freq spacing
f = fmin * df .^ (0:NG-1);   % measurement frequency axis


fe = [0,f,fs/2];
fk = fs*[0:Nfft/2]/Nfft; % fft frequency grid (nonneg freqs)

% Fold cepstrum to reflect non-min-phase zeros inside unit circle
% and compute minimum-phase spectrum
% Since 21-channel Mel-filterbank is employed, we interpolate 158 points with ver
% to get a 200-point symmetric cepstral representation

Smp =  exp( fft( fold( ifft( ( clipdb(log([(10.^(y_mod100(:,50)))'
            10^(-1)*ones(1,158) fliplr((10.^(y_mod100(:,50)))')]),-100) )))));

Ns=101;
Smpp = Smp(1:Ns); % nonnegative-frequency portion
wt = 1 ./ (fk+1); % typical weight fn for audio

wk = 2*pi*fk/fs;

[B,A] = invfreqz(Smpp,wk,NZ,NP,wt);
Hh = freqz(B,A,Ns);
```

```
H = tf(B,A,0.1,'Variable','z^-1');
figure(1);
subplot(211)
pzplot(H);
[PP_100, ZZ_100]=pzmap(H)
subplot(212);
plot(db(abs(Hh)));
```

The function *clipdb* used in this analysis is given in the following:

```
function [clipped] = clipdb(s,cutoff)
% [clipped] = clipdb(s,cutoff)
% Clip magnitude of s at its maximum + cutoff in dB.
% Example: clip(s,-100) makes sure the minimum magnitude
% of s is not more than 100dB below its maximum magnitude.
% If s is zero, nothing is done.

clipped = s;
as = abs(s);
mas = max(as(:));
if mas==0, return; end
if cutoff >= 0, return; end
thresh = mas*10^(cutoff/20); % db to linear
toosmall = find(as < thresh);
clipped = s;
clipped(toosmall) = thresh;
```

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

Bibliography

[1] J. Baker, "The DRAGON system - An overview," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 24–29, February 1975.

[2] F. Jelinek, "Continuous speech recognition by statistical methods," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 532–556, April 1976.

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

[4] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 57, no. 4, pp. 1738–1752, April 1990.

[5] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "The subspace Gaussian mixture model - A structured model for speech recognition," *Computer Speech and Language*, vol. 25, no. 2, pp. 404–439, April 2011.

[6] G. E. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, November 2012.

[7] P. C. Woodland, "Speaker adaptation for continuos density HMMs: A review," in *Proc. ISCA ITRW on Adaptatation Methods for Speech Recognition*, August 2001, pp. 11–19.

[8] A. Chan, E. Gouvea, R. Singh, R. Mosur, R. Rosenfeld, Y. Sun, D. Huggins-Daines, and M. Seltzer, *The Hieroglyphs: Building Speech Applications Using Sphinx and Related Resources*. March 11, Third Draft, 2007.

[9] The HTK Toolkit: http://htk.eng.cam.ac.uk.

[10] The Kaldi Toolkit: http://kaldi.sourceforge.net.

[11] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, "Your word is my command: Google search by voice: A case study," in *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*, 2010, ch. 4, pp. 61–90.

[12] A. Hagen, B. Pellom, and R. Cole, "Children's speech recognition with application to interactive books and tutors," in *Proc. Workshop on Automatic Speech Recognition and Understanding (ASRU)*, November 2003, pp. 186–191.

[13] R. Nisimura, A. Lee, H. Saruwatari, and K. Shikano, "Public speech-oriented guidance system with adult and child discrimination capability," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, May 2004, pp. 433–436.

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

[15] A. Hagen, B. Pellom, and R. Cole, "Highly accurate childrens speech recognition for interactive reading tutors using subword units," *Speech Communication*, vol. 49, no. 12, pp. 861–873, 2007.

[16] S. S. Gray, D. Willett, J. Pinto, J. Lu, P. Maergner, and N. Bodenstab, "Child automatic speech recognition for US English: Child interaction with living-room-electronic-devices," in *Proc. INTERSPEECH, Workshop on Child, Computer and Interaction*, 2014.

[17] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hazen, and L. Hetherington, "JUPITER: A telephone-based conversational interface for weather information," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 1, pp. 85–96, January 2000.

[18] J. Wilpon and C. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, May 1996, pp. 349–352.

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

[20] S. Das, D. Nix, and M. Picheny, "Improvements in children's speech recognition performance," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, May 1998, pp. 433–436.

[21] A. Potaminaos and S. Narayanan, "Robust Recognition of Children Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, November 2003.

[22] S. Eguchi and I. J. Hirsh, "Development of speech sounds in children." *Acta oto-laryngologica. Supplementum*, vol. 257, pp. 1–51, 1969.

[23] R. D. Kent, "Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies," *Journal of Speech and Hearing Research*, vol. 9, pp. 421–447, 1976.

[24] J. L. Gauvain and C. H. Lee, "Maximum a-posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[25] 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, pp. 171–185, 1995.

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

[27] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 357–366, 1995.

[28] S. Ghai, "Addressing Pitch Mismatch for Children's Automatic Speech Recognition," Ph.D. dissertation, Department of EEE, Indian Institute of Technology Guwahati, India, October 2011.

[29] S. Shahnawazuddin and R. Sinha, "Low-memory fast on-line adaptation for acoustically mismatched children's speech recognition," in *Proc. INTERSPEECH*, 2015.

[30] S. Shahnawazuddin, H. Kathania, and R. Sinha, "Enhancing the recognition of children's speech on acoustically mismatched ASR system," in *Proc. TENCON*, 2015.

[31] S. Lee, A. Potamianos, and S. S. Narayanan, "Acoustics of childrens speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, March 1999, selected Research Article.

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

[33] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, "A review of ASR technologies for children's speech," in *Proc. Workshop on Child, Computer and Interaction*, 2009, pp. 7:1–7:8.

[34] P. G. Shivakumar, A. Potamianos, S. Lee, and S. Narayanan, "Improving speech recognition for children using acoustic adaptation and pronunciation modeling," in *Proc. Workshop on Child Computer Interaction*, September 2014.

[35] M. Russell and S. D'Arcy, "Challenges for computer recognition of children's speech," in *Proc. Speech and Language Technologies in Education (SLaTE)*, September 2007.

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

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

[38] H. Singer and S. Sagayama, "Pitch dependent phone modelling for HMM based speech recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 1992, pp. 273–276.

[39] X. Shao and B. Milner, "Pitch prediction from MFCC vectors for speech reconstruction," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2004, pp. 97–100.

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

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

[42] G. Zavaliagkos, R. Schwartz, and J. Makhoul, "Batch, incremental and instantaneous adaptation techniques for speech recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, May 1995, pp. 676–679.

[43] Z. Zhang, S. Furui, and K. Ohtsuki, "On-line incremental speaker adaptation for broadcast news transcription," *Speech Communication*, vol. 37, no. 3-4, pp. 271–281, 2002.

[44] S. M. Ahadi and P. C. Woodland, "Combined Bayesian and predictive techniques for rapid speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 11, pp. 187–206, 1997.

[45] S. M. Ahadi, "Bayesian and predictive techniques for speaker adaptation," Ph.D. dissertation, University of Cambridge, 1996.

[46] W. Huang, Y. Zhang, X. He, and Q. Bao, "Rapid speaker adaptation for embedded large vocabulary dictation system with sparse training materials," in *Proc. International Conference on Audio, Language and Image Processing (ICALIP)*, 2008, pp. 1069–1072.

[47] K. Shinoda and C. H. Lee, "Structural MAP speaker adaptation using hierarchical priors," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, December 1997, pp. 381–388.

[48] M. J. F. Gales and P. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol. 10, pp. 249–264, 1996.

[49] M. J. F. Gales, "The generation and use of regression class trees for MLLR adaptation," Cambridge University Engineering Department, Tech. Rep., 1996.

[50] C. J. Leggetter and P. C. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression," in *Proc. Workshop on Spoken Language Technology (SLT)*, 1995, pp. 110–115.

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

[52] T. J. Hazen and J. R. Glass, "A comparison of novel techniques for instantaneous speaker adaptation," in *Proc. of European Conference on Speech Communication and Technology*, 1997, pp. 2047–2050.

[53] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 8(4), pp. 417–428, July 1999.

[54] R. Kuhn, J. C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, 2000.

[55] I. T. Jolliffe, *Principal Component Analysis*. Springer-Verlag, Berlin, Germany, 1986.

[56] Y. Tsao, S. M. Lee, and L. S. Lee, "Segmental eigenvoice with delicate eigenspace for improved speaker adaptation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 399–411, May 2005.

[57] Y. Jeong and H. S. Kim, "New speaker adaptation method using 2-D PCA." *IEEE Signal Processing Letters*, vol. 17, no. 2, pp. 193–196, 2010.

[58] Y. Jeong, "Speaker adaptation using probabilistic linear discriminant analysis for continuous speech recognition," *IET Letters*, vol. 49, pp. 1641–1643, 2013.

[59] B. Mak, J. T. Kwok, and S. Ho, "Kernel eigenvoice adaptation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 984–992, September 2005.

[60] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, 1984, pp. 4211–4214.

[61] T. Cai and J. Zhu, "A novel method for rapid speaker adaptation based on support speaker weighting," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2005, pp. 993–996.

[62] B. Mak, T. Lai, and R. Hsiao, "Improving reference speaker weighting adaptation by the use of maximum-likelihood reference speakers," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2006.

[63] W. X. Teng, G. Gravier, F. Bimbot, and F. Soufflet, "Speaker adaptation by variable reference model subspace and application to large vocabulary speech recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 4381–4384.

[64] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer New-York, 2010.

[65] X. Xiao, J. Li, E. S. Chng, and H. Li, "Lasso environment model combination for robust speech recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[66] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[67] W. L. Zhang, D. Qu, W. Q. Zhang, and B. C. Li, "Rapid speaker adaptation using compressive sensing," *Speech Communication*, vol. 55, no. 10, pp. 950–963, November 2013.

[68] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, December 1993.

[69] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal on Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 586–597, December 2007.

[70] J. Duchateau, T. Leroy, K. Demuynck, and H. Van hamme, "Fast speaker adaptation using non-negative matrix factorization," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 4269–4272.

[71] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[72] C. J. Lin, "Projected gradient methods for non-negative matrix factorization," *Neural Computation*, vol. 19, no. 10, pp. 2756–2779, October 2007.

[73] E. Gaussier and C. Goutte, "Relation between PLSA and NMF and implications," in *Proc. SIGIR Conference on Research and Development in Information Retrieval*, 2005, pp. 601–602.

[74] S. Hahm, Y. Ohkawa, M. Ito, M. Suzuki, A. Ito, and S. Makino, "Aspect-model-based reference speaker weighting," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2010, pp. 4302–4305.

[75] S. Zhang, P. A. Olsen, and Y. Qin, "Rapid feature space MLLR speaker adaptation with bilinear models," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4452–4455.

[76] X. Zhang, K. Demuynck, and H. V. hamme, "Latent variable speaker adaptation of Gaussian mixture weights and means," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4349–4352.

[77] K. Yu, "Adaptive training for large vocabulary continuous speech recognition," Ph.D. dissertation, University of Cambridge, 2006.

[78] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *The Journal of the Acoustical Society of America*, vol. 55, pp. 1304–1312, June 1974.

[79] L. Lee and R. C. Rose, "Speaker normalisation using efficient frequency warping procedures," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1996, pp. 353–356.

[80] D. Pye and P. C. Woodland., "Experiments in speaker normalization and adaptation for large vocabulary speech recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1997, pp. 1047–1050.

[81] B. Xiang, U. V. Chaudhari, J. Navratil, G. N. Ramaswamy, and R. A. Gopinath, "Short-time Gaussianization for robust speaker verification," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, pp. 681–684.

[82] G. Saon, S. Dharanipragada, and D. Povey, "Feature space Gaussianization," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2004, pp. 329–332.

[83] M. J. F. Gales, X. Liu, K. C. Sim, and K. Yu, "Investigation of acoustic modeling techniques for LVCSR systems," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2005, pp. 849–852.

[84] 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. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, May 1995, pp. 81–84.

[85] Y. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proc.* 27^{th} *Asilomar Conference on Signals, Systems and Computers*, vol. 1, November 1993, pp. 40–44.

[86] S. J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "A method for largescale, l_1-regularized least squares problems with applications in signal processing and statistics," *IEEE Journal on Selected Topics Signal Processing*, vol. 1, no. 4, pp. 606–617, December 2007.

[87] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, pp. 407–499, 2004.

[88] 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*, 2005, pp. 2761–2764.

[89] OMP-Box v10: http://www.cs.technion.ac.il/∼ronrubin/ software.html.

[90] SpasSM: http://www2.imm.dtu.dk/projects/spasm/.

[91] S. Rosset and J. Zhu, "Piecewise linear regularized solution paths," *The Annals of Statistics*, vol. 35, no. 3, pp. 1012–1030, June.

[92] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, New York, USA, 2006.

[93] M. Bacchiani, "Rapid adaptation for mobile speech applications," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[94] S. Ghai and R. Sinha, "Exploring the effect of differences in the acoustic correlates of adults' and children's speech in the context of automatic speech recognition," *EURASIP Journal on Audio, Speech and Music Processing*, vol. 2010, pp. 7:1–7:15, January 2010.

[95] ——, "Pitch adaptive MFCC features for improving children's mismatch ASR," *Internation Journal of Speech Technology*, vol. 18, no. 3, pp. 489–503, September 2015.

[96] ——, "A study on the effect of pitch on LPCC and PLPC features for children's ASR in comparison to MFCC," in *Proc. INTERSPEECH*, 2011, pp. 2589–2592.

[97] ——, "Analyzing pitch robustness of PMVDR and MFCC features for children's speech recognition," in *Proc. Signal Processing and Communications (SPCOM)*, 2010.

[98] J. L. Flanagan, *Speech Analysis, Synthesis, and Perception (2nd Edition)*. New-York: Springer-Verlag, 1972.

[99] A. V. Oppenheim and R. W. Schafer, "Homomorphic analysis of speech," *IEEE Transactions on Audio and Elctroacoustics*, vol. 16, pp. 221–216, June 1968.

[100] J. R. Deller, Jr., J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals (2nd Edition)*. New York: IEEE Press, 2000.

[101] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, vol. 26, no. 4, pp. 283–297, 1998.

[102] G. Stemmer and F. Brugnara, "Integration of heteroscedastic linear discriminant analysis (HLDA) into adaptive training," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, May 2006, pp. 14–19.

[103] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[104] G. Garau and S. Renals, "Combining spectral representations for large-vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 16, no. 3, pp. 508–518, March 2008.

[105] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[106] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Transactions on Information and Systems*, vol. 90, no. 1, pp. 325–333, 2007.

[107] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 2001, pp. 841–844.

[108] T. Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, April 2007, pp. 1249–1252.

[109] G. Garau and S. Renals, "Pitch adaptive features for LVCSR," in *Proc. INTERSPEECH*, 2008, pp. 2402–2405.

[110] P. McLeod, "Fast, Accurate Pitch Detection Tools for Music Analysis," Ph.D. dissertation, University of Otago, Dunedin, New Zealand, May 2008.

[111] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Klein and K. K. Palival, Eds. Elsevier, 1995.

[112] K. Sjölander and J. Beskow, "Wavesurfer - an open source speech tool," in *Proc. INTERSPEECH*, 2000, pp. 464–467.

[113] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[114] R. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 1998, pp. 661–664.

[115] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models." *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[116] M. Brookes, "VOICEBOX: Speech Processing Toolbox for MATLAB," http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 2005.

[117] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech recognition toolkit," in *Proc. Workshop on Automatic Speech Recognition and Understanding (ASRU)*, December 2011.

[118] S. Matsoukas, R. Schwartz, H. Jin, and L. Nguyen, "Practical implementations of speaker-adaptive training," in *Proc. DARPA Speech Recognition Workshop*, 1997.

[119] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to dnn-based children's and adults' speech recognition," in *Proc. Workshop on Spoken Language Technology (SLT)*, December 2014, pp. 135–140.

[120] A. Metallinou and J. Cheng, "Using deep neural networks to improve proficiency assessment for children english language learners," in *Proc. INTERSPEECH*, 2014, pp. 1468–1472.

[121] H. Liao, G. Pundak, O. Siohan, M. K. Carroll, N. Coccaro, Q. Jiang, T. N. Sainath, A. W. Senior, F. Beaufays, and M. Bacchiani, "Large vocabulary automatic speech recognition for children," in *Proc. INTERSPEECH*, 2015, pp. 1611–1615.

[122] S. P. Rath, D. Povey, K. Veselý, and J. Černocký, "Improved feature processing for deep neural networks," in *Proc. INTERSPEECH*, 2013.

[123] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription." in *Proc. Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 24–29.

[124] A. Rahman Mohamed, G. E. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling." in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4273–4276.

[125] A. Mohamed, G. Dahl, and G. E. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Speech and Audio Processing*, vol. 20, no. 1, pp. 14–22, January 2012.

[126] J. Fainberg, "Improving Children's Speech Recognition through Out of Domain Data Augmentation," Master's thesis, School of Informatics University of Edinburgh, 2015.

[127] T. N. Sainath, A. R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 8614–8618.

[128] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 6645–6649.

[129] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *Proc. Workshop on Spoken Language Technology (SLT)*, December 2012, pp. 366–369.

[130] O. Abdel-Hamid and H. Jiang, "Rapid and effective speaker adaptation of convolutional neural network based models for speech recognition," in *Proc. INTERSPEECH*, 2013, pp. 1248–1252.

[131] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 7947–7951.

[132] S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, and Q. Liu, "Fast adaptation of deep neural network based on discriminant codes for speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1713–1725, December 2014.

[133] T. Tan, Y. Qian, and K. Yu, "Cluster adaptive training for deep neural network based acoustic model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 459–468, March 2016.

[134] F. Grezl and P. Fousek, "Optimizing bottle-neck features for lvcsr," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2008, pp. 4729–4732.

[135] F. Grezl, M. Karafiat, and M. Janda, "Study of probabilistic and bottle-neck features in multilingual environment," in *Proc. Workshop on Automatic Speech Recognition and Understanding (ASRU)*, December 2011, pp. 359–364.

[136] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks." in *Proc. INTERSPEECH*, 2011, pp. 237–240.

[137] J. Gehring, Y. Miao, F. Metze, and A. Waibel, "Extracting deep bottleneck features using stacked auto-encoders," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 3377–3381.

[138] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, 2000, pp. 1635–1638.

[139] J. Niu, Y. Qian, and K. Yu, "Acoustic emotion recognition using deep neural network," in *Proc. International Symposium on Chinese Spoken Language Processing (ISCSLP)*, September 2014, pp. 128–132.

[140] M. R. Amer, B. Siddiquie, C. Richey, and A. Divakaran, "Emotion detection in speech using deep networks," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 3724–3728.

[141] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, December 2014.

[142] S. E. Bou-Ghazale and J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 429–442, July 2000.

[143] S. Ramamohan and S. Dandapat, "Sinusoidal model-based analysis and classification of stressed speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 737–746, May 2006.

[144] J. W. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215–1247, September 1993.

[145] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time Signal Processing (2nd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1999.

[146] S. S. Stevens, J. Volkmann, and E. B. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, January 1937.

[147] E. Zwicker, "Subdivision of the audible frequency range into critical bands," *The Journal of the Acoustical Society of America*, vol. 33, no. 2, p. 248, 1961.

[148] Z. Fang, Z. Guoliang, and S. Zhanjiang, "Comparison of different implementations of MFCC," *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, November 2001.

[149] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[150] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 342–350, June 1981.

[151] J. S. Mason and X. Zhang, "Velocity and acceleration features in speaker recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 1991, pp. 3673–3676.

[152] H. Fletcher, "Auditory patterns," *Reviews of Modern Physics*, vol. 11, pp. 47–65, January 1940.

[153] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[154] S. S. Steven, "On the psychopysical law," *Psychological Review*, vol. 64, pp. 153–181, 1957.

[155] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *The Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[156] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[157] 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*, vol. 39, no. 1, pp. 1–38, 1977.

[158] B. H. Juang and L. R. Rabiner, "The segmental k-means algorithm for estimating parameters of hidden Markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 9, pp. 1639–1641, September 1990.

[159] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, A. Rastrow, R. Rose, P. Schwarz, and S. Thomas, "Subspace Gaussian mixture models for speech recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2010, pp. 4330–4333.

[160] R. C. Rose, S.-C. Yin, and Y. Tang, "An investigation of subspace modeling for phonetic and speaker variability in automatic speech recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 4508–4511.

[161] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Speech and Audio Processing*, vol. 16, no. 5, pp. 980–988, July 2008.

[162] S. Haykin, *Neural Networks: A Comprehensive Foundation (2nd Edition)*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998.

[163] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[164] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[165] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, pp. 1771–1800, 2002.

[166] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 215–219.

[167] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. Advances in Neural Information Processing Systems*, 2007, vol. 19, pp. 153–160.

[168] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks." in *Proc. INTERSPEECH*, 2011, pp. 437–440.

[169] J. O. S. III, https://www.dsprelated.com/showcode/20.php.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

List of Publications Related to Thesis

Refereed International Journal Publications

- Accepted manuscripts:

  1. S Shahnawazuddin and Rohit Sinha, *"Improved Bases Selection in Acoustic Model Interpolation for Fast On-Line Adaptation"*, **IEEE Signal Processing Letters**, vol. 21, no. 4, pp. 493-497, April 2014.

- Manuscripts under review:

  1. S Shahnawazuddin and Rohit Sinha, *"Sparse Coding over Redundant Dictionaries for Fast Adaptation of Speech Recognition System"*, **Computer Speech and Language, Elsevier** (third review).
  2. Rohit Sinha and S Shahnawazuddin, *"Assessment of Pitch-Adaptive Front-End Signal Processing for Children's Speech Recognition"*, **Special Issue on Language and Interaction Technologies for Children, Computer Speech and Language, Elsevier**.

Refereed International Conference Publications

- Accepted manuscripts:

1. S Shahnawazuddin, Abhishek Dey and Rohit Sinha, *"Pitch-Adaptive Front-end Features for Robust Children's ASR,"* in Proc. **INTERSPEECH,** 2016.
2. Rohit Sinha, S Shahnawazuddin and Patri Satya Karthik, *"Exploring the Role of Pitch-Adaptive Cepstral Features in Context of Childrens Mismatched ASR,"*, in Proc. **SPCOM,** 2016.
3. S Shahnawazuddin, Hemant Kumar Kathania and Rohit Sinha, *"Enhancing the Recognition of Children's Speech on Acoustically Mismatched ASR System,"* in Proc. **TENCON,** 2015.
4. S Shahnawazuddin and Rohit Sinha, *"Low-memory Fast On-line Adaptation for Acoustically Mismatched Children's Speech Recognition,"* in Proc. **INTERSPEECH,** 2015.
5. S Shahnawazuddin and Rohit Sinha, *"A Low Complexity Model Adaptation Approach involving Sparse Coding over Multiple Dictionaries,"* in Proc. **INTERSPEECH,** 2014.
6. Hemant Kumar Kathania, S Shahnawazuddin and Rohit Sinha, *"Enhancing the Recognition of Children's Speech on Acoustically Mismatched ASR System,"* in Proc. **SPCOM,** 2014.

List of Other Publications

1. S Shahnawazuddin, Deepak Thotappa, Abhishek Dey, Siddika Imani, S R M Prasanna and Rohit Sinha, *"Improvements in IITG Assamese Spoken Query System: Background Noise Suppression and Alternate Acoustic Modeling"*, **Journal of Signal Processing Systems, Springer**, 2016.

2. S Shahnawazuddin, Deepak Thotappa, B D Sarma, A Deka, S R M Prasanna and R Sinha, *"Low Complexity On-Line Adaptation Techniques in Context of Assamese Spoken Query System"*, **Journal of Signal Processing Systems, Springer**, vol. 81, issue 1, pp. 83-97, October 2015.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati