SPECTRAL MAPPING USING ARTIFICIAL NEURAL NETWORKS FOR INTRA-LINGUAL AND CROSS-LINGUAL VOICE CONVERSION

A THESIS

submitted by

SRINIVAS DESAI

for the award of the degree

of

Master Of Science (by Research)

in

Computer Science & Engineering

LANGUAGE TECHNOLOGIES RESEARCH CENTER

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY

HYDERABAD - 500 032, INDIA

APRIL 2010

International Institute of Information Technology

Hyderabad, India

CERTIFICATE

This is to certify that the work contained in this thesis titled **Spectral mapping using Artificial Neural Networks for Intra-lingual and Cross-lingual Voice Conversion** submitted by **Srinivas Desai** for the award of the degree of Master of Science (by Research) in Computer Science & Engineering is a bonafide record of research work carried out by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

_____ _____

Date Mr. Kishore Prahallad

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to Kishore Prahallad, my advisor for his guidance, encouragement and support throughout my duration as an MS student at IIIT-H. I thank him for allowing me to explore the challenging world of speech technology and for always finding the time to discuss the difficulties along the way.

I express my sincere gratitude to Prof. B.Yegnanarayana and Prof. Alan W Black for their valuable advices and inputs during this research.

I am grateful to Prof. Rajeev Sangal, Director of Language Technologies Research Centre, IIIT-H for providing an excellent environment for work with ample facilities and academic freedom.

I thank Guruprasad Sheshadri for all his critical comments on my thesis draft which made me learn how to write better.

I am very grateful for having had the opportunity to study among my colleagues: Sachin Joshi, Santhosh Yuvaraj, Anand Arokia Raj, Satish Chandra Pammi, Gopalakrishna, Vijayaditya, Gautam Mantena, Ramakrishna Raju, Lakshmikanth and Bhaskar. In particular, I thank Venkatesh Keri and Veera Raghavendra - for all the support, fruitful discussions and fun times together.

I am also thankful to the graduate students of IIIT and MSIT who participated in several subjective tests throughout my work.

Needless to mention that without the love and moral support of my father, mother and my sister, this work would not have been possible.

Srinivas Desai

ABSTRACT

Keywords: Voice Conversion, Artificial Neural Networks, Spetral Mapping, Error Correction Network, Cross-Lingual Voice Conversion.

Voice conversion is a process of transforming an utterance of a source speaker so that it is perceived as if spoken by a specified target speaker. Applications of voice conversion include secured transmission, speech-to-speech translation and generating voices for virtual characters/avatars. The process of voice conversion involves transforming acoustic cues such as spectral parameters characterizing the vocal tract, fundamental frequency, prosody etc., pertaining to the identity of a speaker. Spectral parameters representing the vocal tract shape are known to contribute more to the speaker identity and hence there have been efforts to find a better spectral mapping between the source and the target speaker. In this dissertation, we propose an Artificial Neural Network (ANN) based spectral mapping and compare its performance against the state-of-the-art Gaussian Mixture Model (GMM) based mapping. We show that the ANN based voice conversion system performs better than that of GMM based voice conversion system.

A typical requirement for a voice conversion system is to have both the source and target speakers record a same set of utterances, referred to as parallel data. A mapping function obtained on such parallel data can be used to transform spectral characteristics from a source speaker to the target speaker. If either of the speakers change then a new transformation function has to be estimated which requires collection of parallel data. However, it is not always feasible to find parallel utterances for training. The complexity of building training data increases if the language of the source speaker and the target speaker is different, which occurs in the case of cross-lingual voice conversion. To circumvent the need of parallel data and to reduce the complexity in building training

data for a cross-lingual voice conversion system, we propose an algorithm which captures speaker specific characteristics (target speaker) so that there is no need of training data from the source speaker. Such an algorithm needs to be trained on only the target speaker data and hence any arbitrary source speaker could be transformed to the specified target speaker. We show that the proposed algorithm could be used in intra-lingual and cross-lingual voice conversion. Subjective and objective evaluation reveals that the quality of the transformed speech using the proposed approach is intelligible and posses the characteristics of the target speaker.

A set of transformed utterances corresponding to results discussed in this work is available for listening at `http://ravi.iiit.ac.in/~speech/uploads/taslp09_srinivas/`

TABLE OF CONTENTS

ACKNOWLEDGEMENTS **iii**

ABSTRACT **iv**

LIST OF TABLES **ix**

LIST OF FIGURES **xi**

ABBREVIATIONS **xiii**

1 INTRODUCTION TO VOICE CONVERSION **1**

1.1 What is voice conversion? 1

1.1.1 Definition 1

1.1.2 Architecture of a voice conversion system 1

1.2 Issues in voice conversion 4

1.3 Issues addressed in this thesis 6

1.4 Contributions 8

1.5 Organization of thesis 9

2 REVIEW ON VOICE CONVERSION METHODS **10**

2.1 Spectral transformation 12

2.1.1 Intra-lingual voice conversion with parallel data 12

2.1.2 Intra-lingual voice conversion with non-parallel data 18

2.1.3 Cross-lingual voice conversion 23

2.2 Source feature and prosody transformation 24

2.3 Evaluation 28

2.3.1 Objective evaluation 28

2.3.2 Subjective evaluation 29

2.4 Applications 31

2.5 Summary 32

3 VOICE CONVERSION USING ARTIFICIAL NEURAL NETWORKS 34

3.1 Intra-lingual voice conversion 34

3.1.1 Database 34

3.1.2 Feature extraction 35

3.1.3 Alignment of parallel utterances 35

3.1.4 Process of training and testing/conversion 36

3.1.5 Spectral mapping using GMM 37

3.1.6 Spectral mapping using ANN 38

3.1.7 Mapping of excitation features 39

3.1.8 Evaluation criteria for voice conversion 40

3.2 Experiments and results 42

3.2.1 Objective evaluation of a GMM based VC system 42

3.2.2 Objective evaluation of an ANN based VC system 43

3.2.3 Subjective evaluation of GMM and ANN based VC systems 45

3.2.4 Experiment on multiple speaker pairs 46

3.3 Enhancements to voice conversion using ANN 48

3.3.1 Appending deltas 48

3.3.2 A hybrid model 49

3.3.3 Transformation with use of contextual features 51

3.4 Summary 52

4 FRAMEWORK FOR SOURCE-SPEAKER INDEPENDENT VOICE CONVERSION 53

4.1 Many-to-one mapping 53

4.2 Models capturing speaker-specific characteristics 55

4.2.1 Vocal tract length normalization 56

4.2.2 Error correction network 57

4.2.3 Experiments with parallel data 58

4.2.4 Experiments using target speaker's data 59

4.2.5 Experiments on multiple speakers database 61

4.2.6 Application to cross-lingual voice conversion 62
4.3 Summary 63

5 Conclusion and future work **64**
5.1 Conclusion 64
5.2 Limitations 65
5.3 Future work 66

LIST OF TABLES

3.1 Objective evaluation of GMM based VC system for various training parameters where Set 1: SLT to BDL transformation; Set 2: BDL to SLT transformation 42

3.2 MCD obtained on the test set for different architectures of an ANN model. (No. of iterations: 200, Learning Rate: 0.01, Momentum: 0.3) Set 1: SLT to BDL; Set 2: BDL to SLT 44

3.3 Average similarity scores between transformed utterances and the natural utterances of the target speaker. 46

3.4 Results of appending deltas and delta-deltas of MCEPs for (SLT(female) to BDL(male) transformation) 49

4.1 Many to one mapping 54

4.2 Results of source speaker (SLT-female) to target speaker (BDL-male) transformation with training on 40 utterances of source formants to target MCEPs on a parallel database. Here **F** represents Formants, **B** represents Bandwidths, Δ and $\Delta\Delta$ represents delta and delta-delta features computed on **ESPS** features respectively. **UVN** represents unit variance normalization. 59

4.3 Subjective evaluation of voice conversion models built by capturing speaker-specific characteristics 61

4.4 Performance of voice conversion model built by capturing speaker-specific features are provided with MCD scores. Entries in the first column represent source speakers and the entries in the first row represent target speakers. All the experiments are trained on 6 minutes of speech and tested on 59 utterances or approximately 3 minutes of speech. . . . 62

4.5 Subjective results of cross-lingual transformation done using conversion model built by capturing speaker-specific characteristics. 10 utterances from each of Telugu (NK), Hindi (PRA) and Kannada (LV) speakers are transformed into BDL male speaker's voice 62

LIST OF FIGURES

1.1 *Block diagram of various modules involved in a voice conversion system* 2

1.2 *The excitation-filter model of speech production* 2

3.1 *Plot of an utterance recorded by two speakers showing that their durations differ even if the spoken sentence is the same. The spoken sentence is "Will we ever forget it" which has 18 phones "pau w ih l w iy eh v er f er g eh t ih t pau pau" according to the US English phoneset.* . . . 36

3.2 *Plot of an utterance recorded by two speakers showing that their durations match after applying DTW. The spoken sentence is "Will we ever forget it" which has 18 phones "pau w ih l w iy eh v er f er g eh t ih t pau pau" according to the US English phoneset.* 36

3.3 *Training module in voice conversion framework.* 37

3.4 *Testing module in voice conversion framework.* 37

3.5 *An architecture of a four layered ANN with N input and output nodes and M nodes in the hidden layers.* 40

3.6 *MCD scores for ANN, GMM+MLPG and GMM (without MLPG) based VC systems computed as a function of number of utterances used for training. The results for GMM based VC systems are obtained using 64 mixture components.* 44

3.7 ***(a)** - MOS scores for 1: ANN, 2: GMM+MLPG, 3: GMM. **(b)** ABX results for 4: ANN, GMM+MLPG(M->F), 5: ANN, GMM+MLPG(F->M), 6: ANN, GMM(M->F), 7: ANN, GMM(F->M)* 46

3.8 ***(a)** MOS and **(b)** MCD scores for ANN based VC system on 10 different pairs of speakers* 47

3.9 *Block diagram for the hybrid approach* 50

3.10 *MCD scores for the hybrid approach with increasing training data* . 50

3.11 *Graph of MCD as a function of context size for varying number of training utterances for SLT (female) to BDL (male). Context 0 indicates the baseline performance.* 51

4.1 *Plot of LPC spectrum of two male speakers with the original spectrum on the left column and normalized spectrum on the right column.* . . 57

4.2 *Plot of LPC spectrum of two female speakers with the original spectrum on the left column and normalized spectrum on the right column.* . . 57

4.3 *A block diagram of an error correction network* 58

4.4 *A plot of MCD scores obtained between multiple speaker pairs with SLT or BDL as target speakers. The models are built from a training data of 24 minutes and tested on 59 utterances (approximately 3 min).* . . . 60

4.5 *A plot of MCD v/s Data size for different speaker pairs and with SLT or BDL as the target speaker.* 61

ABBREVIATIONS

2-D	-	2 Dimensional
ANN	-	Artificial Neural Networks
CLVC		Cross-Lingual Voice Conversion
DFW	-	Dynamic Frequency Warping
DTW	-	Dynamic Time Warping
EM	-	Expectation Maximization
ESPS	-	Entropic Signal Processing System
FDPSOLA	-	Frequency-Domain Pitch Synchronous Overlap and Add
GMM	-	Gaussian Mixture Models
GV	-	Global Variance
HMM	-	Hidden Markov Models
ILVC	-	Intra-Lingual Voice Conversion
LMR	-	Linear Multivariate Regression
LPC	-	Linear Prediction Coefficients
LSF	-	Line Spectral Frequency
MAP	-	Maximum A Posteriori
MCD	-	Mel Cepstral Distortion
MCEP	-	Mel-cepstral Coefficients
MFCC	-	Mel Frequency Cepstral Coefficients
MLLR	-	Maximum Likelihood Linear Regression
MLPG	-	Maximum Likelihood Parameter Generation
MLSA	-	Mel Log Spectrum Approximation
MOS	-	Mean Opinion Scores
PSOLA	-	Pitch Synchronous Overlap and Add
RBF	-	Radial Basis Function
RMS	-	Root Mean Square

STASC	-	Speaker Transformation Algorithm using Segmental Codebooks
TTS	-	Text-to-Speech
VC	-	Voice Conversion
VFS	-	Vector Field Smoothing
VQ	-	Vector Quantization
VTLN	-	Vocal Tract Length Normaization

CHAPTER 1

INTRODUCTION TO VOICE CONVERSION

1.1 What is voice conversion?

1.1.1 Definition

Speech is a natural medium of communication among human beings. A speech signal carries information including the message that is meant to be conveyed, the identity of a speaker and the background/environment. The characteristics of a speech signal corresponding to the identity of a speaker allow us to differentiate between speakers. An ability to control the identity of a speaker is required in applications such as secured speech transmission and speech-to-speech translation. In secured speech transmission, the identity of a speaker needs to be masked. In speech-to-speech translation, a spoken sentence is translated from one language to another, say from English to French, where it is important that the translated French utterance bears the identity of the English speaker. Other applications with a requirement to control the identity of the speaker include generating voices for virtual characters/ avatars in games. The process of controlling or morphing the identity of a speaker is often referred to as voice conversion. The goal of a Voice Conversion (VC) system is to transform an utterance of a source speaker so that it is perceived as if spoken by a specified target speaker.

1.1.2 Architecture of a voice conversion system

A typical architecture of a VC system is shown in Figure 1.1 and it consists of the following components:

1. A feature extraction module

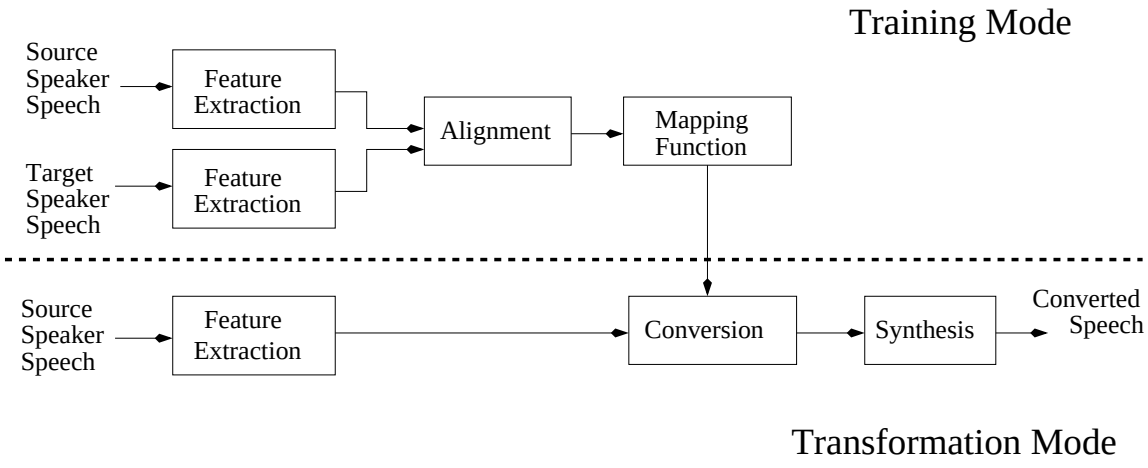

Figure 1.1: *Block diagram of various modules involved in a voice conversion system*

2. Training and transformation modules.

The process of feature extraction assumes a model which is a mathematical representation of the speech production mechanism that makes the analysis, manipulation and transformation of speech signal possible. Gunnar Fant's [1] acoustic theory of speech production postulates that speech production can be modeled by an excitation source and an acoustic filter. Such a model is called as an excitation-filter model and is most widely used in various areas of speech research such as speech synthesis, speech recognition, speech coding, speech enhancement, etc. In this context of excitation-filter modeling, *speech is defined as the output of a time-varying vocal tract system excited by a time-varying excitation signal* [2]. The filter component can be visualized as an acoustic tube with time-varying area function. The input to this filter is an excitation signal which is a mixture of a quasi periodic signal and a noise source. A block diagram of an excitation-filter model is as shown in Figure 1.2.

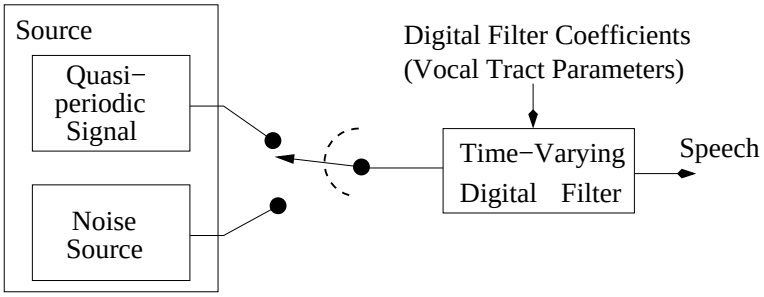

Figure 1.2: *The excitation-filter model of speech production*

Both the excitation and filter are represented by parameters which are usually extracted from the speech signal by performing frame-by-frame analysis, where the size of a frame could vary from 5 ms to 30 ms. Spectral features such as Linear Prediction Coefficients (LPC) [3], Line Spectral Frequencies (LSF) [4], formant frequencies and bandwidths [5], Mel-Frequency Cepstral Coefficients (MFCC) [6], etc., are some of the features generally used to represent the vocal tract shape or the filter. Features such as pitch, residual, glottal closure instants, etc., are used to represent the excitation signal.

Features representing the excitation and filter have cues representing the identity of a speaker, but due to the ease of extraction of the filter features, much emphasis is laid on getting a better spectral transformation. In training phase, standard machine learning techniques such as Vector Quantization (VQ) [7], Hidden Markov Models (HMM) [8] [9] [10], Gaussian Mixture Models (GMM) [11] [12] [13] [14], Artificial Neural Networks (ANN) [15] [16] [17] Dynamic Frequency Warping (DFW) [18] and Unit Selection [19], have been used for learning the transformation from the feature space of the source speaker to that of the target speaker.

Finally, in transformation mode, given a new utterance from the source speaker, the transformation function obtained in the training phase is used to predict the features representing the target speaker. Speech synthesized from these predicted features is perceived as if spoken by the target speaker.

A typical requirement for a VC system is a set of speech utterances recorded by both the source and the target speaker. Depending on the languages in which the training data is available, VC systems can be classified as:

- Intra-Lingual Voice Conversion (ILVC): The source speaker and the target speaker recordings are in the same language.
- Cross-Lingual Voice Conversion (CLVC): The source speaker and the target speaker recordings are in different languages.

1.2 Issues in voice conversion

- **Requirement of parallel data in an ILVC system:** Voice conversion is a process where a source speaker utterance is converted to be perceived as if spoken by a target speaker. A typical requirement for such system is to have both the source and target speakers record a matching set of utterances, referred to as parallel data. A mapping function obtained on such parallel data can be used to transform spectral characteristics from the source speaker to the target speaker [7] [11] [12] [15] [17] [20] [21]. However the use of parallel data has limitations:

    1. If either of the speakers change then a new transformation function has to be estimated which requires collection of parallel data.
    2. The performance of voice conversion is dependent on the match between the source and the target speaker utterances. As the durations of the parallel utterances will typically differ, alignment technique such as dynamic programming is used to have a frame to frame mapping between the utterance pairs. If there are differences between the utterances of source and target speakers in terms of recording conditions, duration, prosody, etc., then it introduces the alignment errors, which in turn leads to poorer estimation of transformation function.
    3. Availability of parallel data is not always feasible. To have a parallel set of recordings from both the speakers and in a naturally time aligned fashion [22] is a costly and time consuming task.

- **Training data for a CLVC system:** The parallel data used for an ILVC system consists of *same set of utterances* recorded by the source and the target speakers. Such parallel data enables us to arrive at a relationship between utterances of the source and the target speakers at a phone/segment level. VC systems built on such data learn a transformation from say phone /a/ of the source speaker to the

phone /a/ of the target speaker. In a CLVC system, as the language of the source and the target speaker is different, there is no possibility of recording the same set of utterances as used in an ILVC system. However, clustering techniques could be used to derive a relationship between features of source and target utterances at phone/segment level. For example, if the utterances of a target speaker could be clustered into K clusters using VQ techniques, then these K clusters could be used to annotate the utterances of the source speaker to derive the relationship between source and target speakers at the cluster level. Such data could be referred to as pseudo-parallel data which could be used to build a CLVC system. There have been various works proposed to exploit the advantage of being able to create pseudo-parallel data for a CLVC system using methods such as: using a speech recognizer [23], a unit selection algorithm [19], class mapping [24], creating pseudo parallel corpus using TTS [10] and adaptation techniques [25]. However, all these methods still need both the speakers data (though not parallel utterances). Hence, there is a need to design an algorithm that captures speaker specific characteristics and uses only the target speaker data. Such an algorithm which needs only the target speaker's data can be used in both ILVC and CLVC frameworks.

- **Smoothing of spectral parameters:** Spectral transformation has always been a concern in this area of VC, hence we find many research works on modeling the conversion function from the source spectral features to the target spectral features. Most state-of-the-art VC systems use GMMs for spectral transformation [11] [12] [13] [14] which needs Maximum Likelihood Parameter Generation (MLPG) [26] to perform a parameter smoothing. However excessive smoothing causes decreased similarity to the target speaker voice [27]. Hence, we intend to find a transformation method that does not need this smoothing.

- **Prosodic conversion:** Prosodic conversion refers to the transformation of prosodic characteristics such as mean fundamental frequency, phone duration, loudness

etc., of the source speaker to that of the target speaker. Most of the current VC methods, use a normalized linear transformation of pitch/fundamental frequency due to which the prosody of the source speaker is copied. i.e, the transformed speech may bear the identity of the target speaker but possess the prosodic characteristics of the source speaker.

- **Quality vs. Similarity:** There is generally a trade-off between the quality of a transformed speech signal and similarity to the target speaker's voice. For some applications, distortion in the transformed speech signal might be tolerated to increase their similarity to the target voice. For example, in singing voice transformations, part of the distortion in the voice conversion output becomes inaudible when mixed with music. Therefore, the similarity versus quality trade-off needs to be maintained in voice conversion algorithm.

1.3 Issues addressed in this thesis

There are two issues in particular that we propose to address in this thesis.

1. To find a better spectral transformation method with the use of parallel training data and hence compare the proposed approach with the state-of-the-art GMM based VC system.

2. To circumvent the requirement of parallel data in intra-lingual voice conversion and to reduce the complexity in obtaining training data for a cross-lingual voice conversion system, we propose an algorithm which captures speaker specific characteristics of the target speaker. Such an algorithm needs to be trained on only the target speaker data and hence any arbitrary source speaker's speech could be transformed to the specified target speaker.

Vocal tract shape between two speakers is non linear and hence ANN based spectral transformation was proposed as this can perform non-linear mapping [15]. Narendranath et. al. [15] used ANNs to transform the source speaker formants to target speaker formants. Results were provided showing that the formant contour of the target speaker can be obtained using ANN. A formant vocoder was used to synthesize the transformed speech, however, no objective or subjective measures were given as to how good the transformed speech was. The use of radial basis function neural network for voice transformation was proposed in [16] [28]. All the above referred methods in [15] [16] [28] need carefully prepared training data which involves manual selection of vowels or syllable regions from both the source and the target speaker. This is a tedious task to make sure that the source and the target features are aligned correctly. Hence, there is a need for an algorithm that does not need any manual selection for training data. The work in [29] also uses ANN for spectral and prosodic mapping, but it is not clear how the proposed ANN based VC system compares with most widely used GMM based VC systems.

We also propose the use of ANNs for spectral mapping and our work differs from earlier approaches in the following ways:

- The proposed approach using ANNs makes use of the parallel set of utterances provided from source and target speakers to automatically extract the relevant training data for mapping of source speaker's spectral features onto the target speaker's acoustic space. Thus our approach avoids any requirement of manual or careful preparation of data.

- Subjective and objective measures are conducted to evaluate the usefulness of ANNs for voice conversion.

- A comparative study between ANN and GMM based VC systems is performed and we show that ANN based VC performs as good as that of GMM based conversion.

- We propose additional techniques that improve the transformation performance. These techniques include use of delta features and use of 2-D features.
- To address the issue of obtaining training data from the source and the target speaker for a CLVC system, we propose an algorithm that captures only the target speaker characteristics and hence does not require the source speaker data at the training stage. In this way, we will be able to transform any arbitrary source speaker to a particular target speaker. The proposed approach is useful for both ILVC and CLVC.

1.4 Contributions

The contributions of this study could be summarized as follows:

- Use of ANNs for VC on continuous speech data without any need for frame selection either manually or computationally.
- Comparison of ANN and GMM for spectral transformation in VC.
- Development of a hybrid framework that combines both ANN and GMM for spectral transformation in VC.
- Development of a novel framework for spectral transformation which captures speaker specific characteristics and hence the training for VC could be done without any need for recordings from a source speaker. i.e, the models are built on only the target speaker data and hence we can transform any arbitrary source speaker onto the trained target speaker's acoustic space.
- Introduction of an error-correction module to improve the performance of the voice conversion system
- Application of the proposed framework in CLVC.

1.5 Organization of thesis

The rest of the thesis is organized as follows.

In **Chapter 2**, a brief explanation is given on various techniques that were proposed for ILVC and CLVC based systems are explained in brief. The issues that still remain unresolved are noted.

Chapter 3 is devoted to the design of a baseline system which explains in detail, a VC system trained on parallel data using both ANN and GMM. A comparison of these two techniques is also done using different speaker pairs and on varying the amount of training data to finally conclude that use of ANNs is better for spectral transformation. Three techniques applied to enhance the performance of a VC system based on ANN are also described in detail.

Chapter 4 proposes a new algorithm that captures speaker specific characteristics and hence resolve the issue of using parallel data for VC training. In the process of designing this algorithm, a new module called error-correction network is proposed which improves the performance of the above mentioned algorithm. Finally, we conclude this chapter with experiments and results of this algorithm when tested in a CLVC scenario.

Finally in **Chapter 5**, the conclusions that can be drawn from this thesis are outlines and some possible research lines for the future are proposed.

A set of transformed utterances corresponding to results discussed in this work is available for listening at `http://ravi.iiit.ac.in/~speech/uploads/taslp09_srinivas/`

CHAPTER 2

REVIEW ON VOICE CONVERSION METHODS

The objective of a Voice Conversion (VC) system is to transform the identity of a source speaker so that it is perceived as if spoken by a specified target speaker. Hence a VC system should be capable of transforming cues representing the identity of a speaker. Studies concerning inter-speaker variations have revealed that there are several parameters in a speech signal, both at the linguistic and at the acoustic level, which contribute to inter-speaker variability and identity of a speaker [30]. Linguistic cues include the language of the speaker, the dialect, choice of lexical patterns, choice of syntactic constructs and the semantic context. The acoustic level features are divided into the segmental and suprasegmental levels.

- Segmental cues depend on the physiological and physical properties of the speech organs which describe the timbre of a speaker's voice. When describing the human voice, people generally refer to the overall quality of a voice as its timbre [10]. The timbre enables the listener to distinguish between different speakers, even when they utter the same text. Timbre is a perceptual attribute, influenced by multiple factors. Acoustic descriptors of timbre include pitch, glottal spectrum and short-time spectrum of the speech signal.

- Suprasegmental cues, on the contrary, are influenced by psychological and social factors and describe the prosodic features related to the style of speaking. They are mainly encoded in pitch, duration and energy contours. These cues are generally obtained by analyzing segments of speech of duration more than 20 ms.

Research has been done in understanding the cues in a spoken utterance that best represent speaker characteristics. Authors of [31] investigated the contribution of F_0,

formant frequencies, spectral envelope and other acoustics parameters towards speaker individuality. They observed that F_0 was the most important feature followed by F_0 intonation pattern and then the spectral tilt. However it was shown that spectral envelope had the greatest influence on speaker individuality, followed by F_0 in [32]. The work in [33] concluded that "it could not be assumed that any single acoustic feature alone could carry the entire individuality information, as a voice/speech is an amalgam of many parameters and the degree or order of importance among the features differ from speaker to speaker".

Although a complete voice conversion system should transform all types of speaker-dependent characteristics of speech, current voice conversion systems are focused only on the acoustic features of voice i.e., fundamental frequency F_0 and spectral characteristics. A majority of them focus on the spectral transformation due to ease of extraction of spectral features from the speech signal. Depending on the languages in which the training data is available, voice conversion systems can be classified as intra-lingual voice conversion and cross-lingual voice conversion. In an intra-lingual voice conversion system, the source speaker and the target speaker record the training utterances in the same language. This system can be further divided into two types: 1) The source and the target speakers record the same utterances (parallel data). 2) The source and the target speaker record different utterances but in the same language (non-parallel data). More information about this topic is given in Section 2.1.1 and 2.1.2 respectively. In a cross-lingual voice conversion system, the source and the target speakers record utterances in two different languages. Various approaches proposed in this context are described in Section 2.1.3. Section 2.2 describes the various methods adopted to transform excitation features and prosody. The subjective and objective evaluation methods are explained in Section 2.3. The applications of voice conversion are discussed in Section 2.4. Section 2.5 ends with a summary and issues which are yet to be addressed.

2.1 Spectral transformation

Referring to the state-of-the-art VC systems, we observe that spectral transformation plays a vital role in VC. Hence, this section is dedicated to understanding various approaches designed to perform spectral transformation efficiently.

2.1.1 Intra-lingual voice conversion with parallel data

Codebook mapping for spectral transformation

An earlier attempt for Intra-Lingual Voice Conversion(ILVC) proposed the use of codebook based transformation method [7]. The basic idea of this technique was to make mapping codebooks which represent the correspondence between the source and the target speaker pair. In order to generate a codebook, all utterances recorded by the two speakers are vector quantized (also called hard clustering) frame-by-frame. As the durations of the parallel utterances will typically differ, Dynamic Time Warping (DTW) was used to align the utterances. Hence, a frame-to-frame correspondence between the source and the target speaker codebook entries (i.e, codewords) were obtained. To ensure that the transformation was not biased due to an unequal number of vectors in each cluster, a weighting function was estimated based on the count of the number of vectors in each cluster. The mapping thus obtained was used to transform a source speaker's speech to be perceived as that of the target speaker. However, a vector quantizer clusters data into discrete sets and hence it causes discontinuities and reduces the quality of converted speech. To reduce these discontinuities, fuzzy vector quantization [34] and weighted vector quantization [33] were proposed.

An algorithm called Speaker Transformation Algorithm using Segmental Codebooks (STASC) was proposed, which was a modification of the codebook based mapping technique [35]. In this approach, a left-to-right Hidden Markov Model (HMM) with no skip state was trained for each utterance of the source speaker. The HMM was

initialized using segmental K-means algorithm and trained using Baum-Welch algorithm. For every 40 milliseconds, a new state was added to the HMM topology. Hence, the number of states for each utterance was directly proportional to the duration of the utterance. The utterances of the target speaker were force-aligned with corresponding source utterance HMM using viterbi algorithm. The training of HMMs and alignment was performed on MFCCs which represent spectral characteristics. This automatic alignment procedure was called sentence-HMM method. After sentence-HMM based alignment, LSF vectors, fundamental frequency values, durations and energy values were calculated in the corresponding source and target HMM states. The state arithmetic means of these acoustic features were computed and stored in source and target speaker codebooks. The mapping thus obtained after the alignment was used for transformation. However, a sentence-HMM based alignment was not found to be robust when there were differences in the prosody, accent or recording conditions. This was because speaker dependent HMMs were built on the source speaker's speech and were used to segment the target speaker's speech. Since the acoustic properties of the source speaker and the target speaker may not be same, the above method would lead to alignment mismatches hence leading to distortion in the output speech.

Authors of [27] proposed a method using phonetic-HMM which gave better alignment than the sentence-HMM based method. The proposed idea was to force-align the corresponding phoneme sequences, rather than between the whole sentences. The phonetic context of each aligned acoustic feature pair was determined from the labels. The acoustic features extracted were paired on a frame-by-frame basis using the alignment information. It was concluded that the proposed method performed better than the sentence-HMM based method and that it could be slightly modified to fit the needs of a cross-lingual voice conversion.

Two approaches for learning spectral conversion methods, namely Dynamic Frequency Warping (DFW) and Linear Multivariate Regression (LMR) were proposed and compared in [18]. In this method, the source and the target speakers recorded the same

set of utterances which were aligned using DTW. The utterances of the source speaker were partitioned into Q non-overlapping classes (referred to as acoustic classes) by means of vector quantization. Let $C^{s,q}$ denote a set of source spectral vectors belonging to q^{th} class where $q = 1, 2, ..., Q$ and let M_q denote the total number of vectors in the q^{th} class. The source spectral vectors and the target spectral vectors were aligned (mapped), hence dividing the source spectral vectors into Q classes also divides the target spectral vectors automatically.

LMR maps each acoustic class of source speaker to the corresponding class of the target speaker using a linear transformation. The linear regression transformation matrix P_q that minimizes the mean-squared error between aligned source and target vectors is obtained as a solution to the following minimization problem:

$$\arg\min_{P_q} \sum_{k=1}^{M_q} ||C_k^{t,q} - P_q C_k^{s,q}||^2 \tag{2.1}$$

LMR modifies the spectrum shape of a source speaker's utterance to match the target speakers utterance spectrum.

DFW represents the correspondence between the source frequency axis and the target frequency axis by a warping function $\alpha(w)$ such that

$$Y(w) = X(\alpha(w)) \tag{2.2}$$

where Y(w) and X(w) denote the target and source utterance power spectrum respectively. A warping function was determined for each acoustic class/cluster of source-target spectrum pairs. The final warping function was defined as the median warping in each class. In DFW only formant positions were moved but their amplitudes could not be modified, which would not lead to an effective transformation of the speaker. Hence, the authors concluded that LMR performs better than DFW with respect to transforming voice quality, but produces some audible distortions.

GMMs for spectral transformation

It was the development of continuous probabilistic modeling and transformation that lead to a considerable improvement in voice conversion performance. Use of GMMs to model and transform the source and target features was proposed in [12]. Clustering using GMMs was called soft clustering as they provide continuous models. Conversion of spectral envelopes using GMMs have been demonstrated to be more robust and efficient than the transformation based on VQ. However, another method that models the joint distribution of source and target features using GMM was proposed in [36] to improve the transformation performance. This method predicts the target spectral features from the source spectral features. The authors were able to conclude that such a method improves the speaker recognizability over the conventional GMM based method proposed in [12]. A common problem shared by all spectral envelope conversion methods is the broadening of the spectral peaks, expansion of the formant bandwidths and over-smoothing caused by the averaging effect of parameter interpolation [37]. This phenomenon makes the converted speech sound slightly muffled. As a solution to this, Maximum Likelihood (ML) transformation approach was proposed in [38] [39], which estimates ML taking into account the global variance of the converted spectra in each utterance and reduces the over-smoothing problem. Another technique proposed to resolve the issue of degradation of spectral transformation was the use of Dynamic Frequency Warping (DFW) [39]. However, used in a framework of a Text-to-Speech (TTS) system where VC was treated as a post-processing block after the TTS to generate new voices.

Sub-band processing for spectral transformation

The role of different factors on the perception of speaker identity was investigated in [40]. Four acoustic features were considered for this study. It was observed that the features representing the vocal tract system were more important followed by both F_0 and duration. However F_0 was considered to be more important than duration, only when

the transformation was of a cross-gender type such as a male-to-female speaker transformation or a female-to-male speaker transformation. The least important feature was found to be energy contour. These conclusions were the results of a subjective test conducted in [40]. They also found that the range of 1 kHz to 2.7 kHz in speech spectrum was the most important for speaker identity. Hence, they proposed two new methods based on sub-band processing using discrete wavelet transforms. The proposed framework had the flexibility to analyze different frequency bands using different amounts of spectral resolution.

ANNs for spectral transformation

Models for VC based on Artificial Neural Networks (ANN) were proposed in [15] based on the property that a multi layered feed-forward neural network using non-linear processing elements could capture an arbitrary input-output mapping. This generalization property of ANN helps in the faithful transformation of formants across speakers, avoiding the use of large codebooks. The training scheme of the conversion system based on ANN in [15] consisted of formant analysis, followed by a learning phase in which the implicit formant conversion between the source and target speaker utterances for the first three formants was captured by a neural network. In the transformation phase, the three formants extracted from each frame of the source speaker's speech were given as input to the trained ANN to obtain converted formants. The converted formants, together with the source pitch contour modified to suit the average pitch of the target speaker, were used in a formant vocoder to synthesize speech with the desired vocal tract system characteristics. Radial Basis Function (RBF) networks with Gaussian basis function was proposed in [16]. RBFs were introduced in the referred study to compensate for the effect of training time and complexity in back-propagation algorithm.

HMMs for spectral transformation

A conventional VQ based method uses only static information to determine the clusters. However, as the dynamic characteristics are also important for this, HMM based VC was proposed in [8]. In this approach, the goal was to use delta parameters or transition probabilities to improve the efficiency of transformation. In the training stage, given an utterance and its equivalent transcript, a HMM was trained to find an optimal state sequence on the source speaker data. Each state of the source speaker data was quantized and the output was called as recognition codebook. As the proposed HMM has vector quantized states, it was called a Hidden Markov VQ Model (HMVQM). All the target speaker data corresponding to each codeword of the source speaker data was collected and their means were calculated. This set of means on the target speaker data was called synthesis codebook. The obtained one-to-one mapping between the codewords of each state was used for conversion. Since the proposed method performs mapping at state level and models dynamic characteristics as transition probabilities, they were able to conclude that such an approach was much better than the conventional VQ based method.

A similar segmentation system based on HMM was proposed in [9], but the transformation function associated with each state was based on Maximum Likelihood Linear Regression (MLLR). MLLR is a model adaptation technique that estimates a set of linear transformations for the mean and variance parameters of a Gaussian mixture HMM system. If V_t represents target vector sequence and V_s represents the source vector sequence then

$$V_t = AV_s + b \tag{2.3}$$

The parameters A and b are to be determined such that the target data gives maximum likelihood. The effect of these transformations shifts the component means and alter the variances in the initial system so that each state in the HMM system is more

likely to generate the target data. During synthesis, the parameters A and b are used to transform the source speech.

2.1.2 Intra-lingual voice conversion with non-parallel data

Most of the approaches described above, use parallel corpora for training i.e, the source speaker and the target speaker record the same set of utterances. Since it is not always feasible to find parallel utterances for training, there were methods that were proposed with the goal of reducing the recordings from the source speaker. All such methods use non-parallel training data, the goal of which is to find a one-to-one correspondence between the frames of source and target speaker. The different kinds of methods that work with the non-parallel data can be classified as follows:

- Class mapping
- Speech recognition
- Pseudo parallel corpora created for a TTS
- Unit selection
- Adaptation techniques

Class mapping

In this method, the source speaker data and the target speaker data were vector quantized to obtain K_s and K_t number of clusters [24] respectively. For each centroid in source class, a similar centroid in target class has to be estimated. Since the clusters obtained were from two different speakers, a distance measure to estimate the similarity measure between the clusters could not be used directly. Hence, Dynamic Frequency Warping (DFW) was used to compensate for the speaker characteristics and hence bring

the centroids to a common domain. Finally, the vectors inside each class were mean-normalized and the frame-level alignment was performed by finding the nearest neighbor for each source vector in the corresponding target class. This technique was evaluated using objective measures and it was found that the performance of this method was not as good as that which could be obtained using parallel data. However this method was proposed as a starting point for further improvements that lead to the development of unit selection based VC.

Speech recognition

In [23], a speech recognizer based on speaker-independent HMMs was used to label all the source and target frames with a state index. Given the state sequence of one speaker, the alignment procedure consists of finding the longest matching state sub-sequences from the other speaker until all the frames were paired. The HMMs used for this task was observed to be valid for intra-lingual alignment. However, the suitability of such models for cross-lingual alignment tasks was not tested.

Pseudo-parallel corpus created by a TTS

In this technique, the utterances recorded by the source speaker were used to build a TTS. All utterances recorded by the target speaker were synthesized using a TTS built on source speaker data. The synthesized utterances together with the target data form parallel data on which algorithms explained in Section 2.1.1 could be applied. However, this solution could be put into practice only under certain conditions:

- The TTS system uses linguistic knowledge to generate artificial sentences, so the language of the desired output sentence has to be the same as the language of the recorded units. Therefore, this kind of technique was restricted to intra-lingual context.
- The size of the training corpus has to be large enough to build a TTS system.

If only a few minutes of audio were available for building the TTS for source speaker, the resulting low-quality synthetic speech leads to a distorted conversion function that introduces artifacts into the converted speech.

Such a system was used in [10] where the spectral transformation performance using machine learning techniques such as GMM, HMM and CART were compared. The overall performance of the proposed intra-lingual system (using CART) was found to be comparable with the performance of other intra-lingual systems (using GMM/HMM) trained under parallel conditions. The evaluation results confirm that using a TTS for non-parallel alignment leads to satisfactory results. As the proposed algorithm performed satisfactorily in case of ILVC, the validity of the algorithm was also tested in CLVC framework. The similarity scores obtained by the cross-lingual system were slightly lower than those of the intra-lingual system. This was probably due to the fact that Spanish and English used in [10] have different phoneme sets. Consequently, the transformation functions trained for one of these languages were not capable of converting the phonemes of the other language with the same accuracy.

Unit selection paradigm

The proposed algorithm in [19] [41] [42] was to compare two different databases and find frames in the source database that were nearest to any of the frames in the target speaker database. The distance measure was computed by a cost function such as the one used in TTS systems to concatenate two units. In a unit selection based TTS system, there are two costs that are involved: target cost and concatenation cost. Minimizing these costs ensures that the distance between the source and target features are minimized and a maximum continuity is obtained between the selected units. However, a disadvantage in this technique when using a large training database was that the vectors that occur initially will most likely be repeated and hence, the vectors at the bottom may never be selected. This may cause degradation in output quality. In order to achieve a better performance, all the training vectors should take part in the alignment, so that no

phonetic areas are left uncovered in the acoustic spaces of the speakers.

Therefore a new method for estimating pseudo-parallel data was proposed in [30]. A nearest neighbor of each source vector in the target acoustic space and the nearest neighbor of each target vector in the source acoustic space allowing one-to-many and many-to-one alignments were mapped. When a VC system using GMM was trained on such aligned data it was observed that an intermediate converted voice was obtained, i.e, it was neither recognized as a source speaker's voice nor as the target speaker's voice. However the proposed approach was applied on the transformed data and target speaker data which resulted in an output closer to the target speaker than the previous transformed sentences. If this procedure was followed iteratively, the final voice was found to converge to the target speaker.

Adaptation techniques

The technique proposed in [43] for voice conversion was based on building a transformation module on the existing parallel data of an arbitrary source-target speaker pair and adapt this model to the particular pair of speakers for which no parallel data was available. Suppose A and B are the two speakers between whom we need to build a transformation function, but the recorded utterances by these speakers are not parallel. Suppose that we also have parallel recorded utterances from speakers C and D. We could then estimate a transformation function between speakers C and D and use adaptation techniques to adapt the conversion model to the speakers A and B.

In [43], the spectral vectors that correspond to the source speaker of the parallel corpus were considered as realizations of random vector x, while y corresponds to the target speaker of the parallel corpus. From the non-parallel corpus, x' is considered as realization of random vector for the source speaker and y' for the target speaker. An attempt was then made to relate the random variables x and x', as well as y and y', in order to derive a conversion function for the nonparallel corpus based on the parallel corpus parameters. An assumption made is that x is related to x' by a probabilistic

linear transformation, as shown in the equation below.

$$x' = \begin{cases} A_1 x + b_1 & \text{with probability } p(\lambda_1|\omega_i) \\ A_2 x + b_2 & \text{with probability } p(\lambda_2|\omega_i) \\ . & . \\ . & . \\ A_N x + b_N & \text{with probability } p(\lambda_N|\omega_i) \end{cases} \tag{2.4}$$

where

$$\sum_{j=1}^{N} p(\lambda_j|\omega_i) = 1,\ i = 1,2,....M. \tag{2.5}$$

In the above equation M is the number of mixtures of the GMM corresponding to the joint vector sequence of the parallel data. y and y' are also related to each other by a probabilistic linear transformation as shown in the equation below.

$$y' = \begin{cases} C_1 x + d_1 & \text{with probability } p(\hat{\lambda_1}|\omega_i) \\ C_2 x + d_2 & \text{with probability } p(\hat{\lambda_2}|\omega_i); \\ . & .; \\ . & .; \\ C_N x + d_N & \text{with probability } p(\hat{\lambda_N}|\omega_i) \end{cases} \tag{2.6}$$

where

$$\sum_{p=1}^{L} p(\hat{\lambda_p}|\omega_i) = 1,\ i = 1,2,....M. \tag{2.7}$$

The unknown parameters i.e, the matrices A_j, C_p and the vectors b_j, d_p were estimated from the non-parallel data by applying EM algorithm. Therefore x' and y' would

be estimated as a linearly constrained maximum likelihood of the GMM parameters. The issue with such an algorithm was that a parallel database was needed on which the initial model could be estimated.

2.1.3 Cross-lingual voice conversion

Cross-Lingual Voice Conversion (CLVC) is the most extreme case of intra-lingual voice conversion with no parallel data where both the training utterances and the the training languages are different. Voice conversion systems dealing with different languages have some special requirements because the utterances available for training are characterized by different phoneme sets. The different categories of approaches for CLVC are:

- Class Mapping
- Unit selection

Class Mapping

This method of class mapping proposed in [27] was a modification of the STASC method proposed in [35] described in Section 2.1.1. The training stage in [27] starts with the extraction of acoustic parameters from the source and the target speaker training recordings. The vocal tract characteristics were represented in two forms: Mel Frequency Cepstral Coefficients (MFCCs) for the alignment stage and line spectral frequencies (LSFs) for the transformation stage. All the source and target recordings were segmented using phonetic-HMM based segmentation. The segmentation could be performed in two ways: text-independent and text-dependent. For non-parallel training databases, all parameters and information extracted on their phonetic context were saved. The mapping of the source and the target acoustic spaces was performed in the transformation stage by context matching. The advantage with this method over the

STASC method was that it was designed to be independent of text and hence could be trained on non-parallel training utterances. The vocal tract transformation function was estimated directly from the speech frames in the source and target training database instead of state averages as in STASC. This helps to perform more detailed vocal tract transformation.

Unit selection

Given a set of source and target non-parallel speech utterances, the goal was to find frames that phonetically correspond to each other. As the text of the spoken utterances was not used, K-means algorithm was applied to divide the spectral space into clusters to represent a set of artificial phonetic classes. Both the source and the target speaker data was clustered and a mapping between the clusters had to be estimated. But as the mapping was obtained between features of two different speakers, the training data had to be normalized. This normalization was done by using VTLN [44] which was a technique used to compensate for the effect of speaker-dependent vocal tract lengths. For every target cluster, a nearest source cluster was estimated, i.e, a mapping between the phonetic classes of source and target speech was estimated. This class mapping was extended to find a frame-to-frame mapping in the source-target cluster pairs.

Authors in [30] proposed an iterative method of aligning speech frames which was also explained in Section 2.1.2.4. It was observed that for a CLVC case they were able to get an acceptable level of performance.

2.2 Source feature and prosody transformation

Though the residual signal is impulse-like for voiced frames and noise-like for unvoiced frames, it contains the glottal characteristics that are not modeled by spectral features. The excitation signal also contains information that could help to achieve the required conversion performance and quality.

An approach towards ILVC [7] (explained in Section 2.1.1), assumes an excitation-filter model of speech and hence to obtain a good performance on voice conversion, features of both the excitation and filter were transformed. Pitch frequencies and power values were quantized, similar to the technique used to quantize the spectral features. A mapping function thus learned on this quantized data was used for transformation. However, the use of a vector quantizer causes discontinuities and reduces the quality of the transformed speech.

PSOLA was used to modify the source speaker's excitation signal to match that of the original target speaker's excitation signal in [18]. This modification was done on both the time and pitch scale, resulting in a better output. However, this experiment was feasible only as a study to understand how well the spectral feature transformation performs because we do not have access to the target speakers speech and hence its excitation in real time.

Based on the idea that a speech signal contains non-linearities mainly present in the residual signal, residuals have been modeled by a long-delay nonlinear predictor using a time-delay neural network [45]. Once the predictor was estimated, a mapping codebook was built to transform the residual signal. It was reported that the naturalness of the converted speech increases when introducing the residual mapping, but some buzzy quality or click noises appear in regions with mixed voicing. STASC [35] deals with residual signals too. An excitation transformation filter was formulated for each codeword entry, using the excitation spectra of the source speaker and the target speaker, in the same way that the vocal tract conversion filter was built.

A different approach from residual mapping was proposed in [11], where the residual signal was predicted from the vocal tract parameters, instead of transforming the source speaker residual. The underlying assumption of the proposed approach was that for a particular speaker and within some phonetically similar class of voiced speech, the residuals were similar and predictable. For each class of phonetically similar units, residual codebooks were stored. Use of such residual codebooks was to produce an

output nearly indistinguishable from the original speech. Recently, the same prediction strategy has been adopted by other researchers [46] [47].

The authors of [29] have also reported the use of residual prediction to transform a source speaker's speech. ANNs were used to train a mapping network between the source speaker's residual and the corresponding target speaker's residual signal. The authors have also reported experiments which include mapping of intonation, duration and energy patterns. Two ANNs were trained for intonation mapping to capture (i) gross level variations which depend on semantics of speech, (ii) finer level variations which indicate prominence of individual words. Two ANNs were trained for duration mapping to capture (i) duration patterns of a syllable, (ii) duration pattern of non-speech/pause. Two ANNs were trained for energy pattern mapping to capture (i) energy at utterance level and (ii) energy at syllable level. Finally the authors were able to conclude that using different ANNs to capture prosodic variations were helpful in transforming a source speaker's utterance to the target speaker.

A logarithm Gaussian normalized transformation [48] was used to transform the source speaker F_0 to target speaker F_0 as indicated in the equation (1) below. The assumption in this case was that the major cues of speaker identity lie in the spectral features and hence just a linear transformation was sufficient to transform excitation characteristics.

$$\log(F_{0\,conv}) = \mu_{tgt} + \frac{\sigma_{tgt}}{\sigma_{src}}(\log(F_{0\,src}) - \mu_{src}) \tag{2.8}$$

where μ_{src} and σ_{src} are the mean and variance of the fundamental frequencies in logarithm for the source speaker, $F_{0\,src}$ is the pitch of source speaker and $F_{0\,conv}$ is the converted pitch frequency.

Prosodic conversion refers to the transformation of the prosodic characteristic of a source speaker (mean of the fundamental frequency, phoneme duration, loudness) to the prosodic characteristics of a target speaker. Prosody transformation is beyond the scope

of this thesis. However, some relevant works will be referred to in order to complete the discussion on state-of-the-art VC systems. Prosodic conversion is the aspect less studied in VC systems. On one hand, most approaches described for spectrum conversion only scale the pitch of the source speaker to resemble that of the target one, without dealing with phoneme durations. On the other hand, there are some approaches that construct a prosodic conversion system similar to the spectrum mapping, such as STASC [35], [49] or MLLR [50]. The method proposed in [51] consists of a stochastic system that transforms pitch contours taking into account multiple pitch parameters, such as: pitch, pitch declination and variances, according to the length of the utterances. The basic idea of this system was to model pitch evolutions of a phrase by a declination line plus a normal distribution to take into account the variation of the pitch around that line.

Recently, prosodic conversion has been studied in the framework of speech-to-speech translation in order to improve the quality of the output prosody. The authors of [52] proposed the use of the intonation of the speaker of the source language to improve the quality of the intonation of the target language. To take into account the converted prosody, the following speech generation process was proposed. First, the prosodic features of the source speaker were estimated. Second, a prosodic mapping module performed the transformation of the estimated features in order to enrich the output of the translation module. Finally, the speech synthesis module produced the output waveform signal using prosody generated by the prosody generation module, which takes advantage of the enriched text.

Whilst state-of-the-art implementations are capable of achieving reasonable conversions between speakers with similar voice characteristics and prosodic patterns, they do not work as well in scenarios where the differences between the source and the target speech are more extreme. This was mainly due to limitations in the modeling and conversion of the voice source and prosody. Hence, in [37], a refined modeling and transformation of the voice source and duration was proposed to increase the robustness of voice conversion systems in extreme applications. In addition, the developed tech-

niques were tested in a speech repair framework. Voice source modeling refinement involved the use of Liljencrants-Fant model instead of the linear prediction residuals employed by the existing implementations to represent the voice source. A speech model was also developed which automatically estimates voice source and vocal tract filter parameterizations. The use of this speech modeling technique for the analysis, modification and synthesis of speech allows the application of linear transformations to convert voice source parameters. The performance of the developed conversion system has been shown to be comparable to that of state-of-the-art implementations in terms of speaker identity, but to produce converted speech with a better quality. Regarding duration, a decision tree approach was proposed to convert duration contours. Its application has been shown to reduce the mean square error distance between the converted and target duration patterns and to increase their correlation.

2.3 Evaluation

Transformation performance in voice conversion systems is generally evaluated using both objective and subjective measures. Objective evaluations are indicative of conversion performance and could be useful to compare different algorithms within a particular framework. However, objective measures on their own are not reliable, since they may not be directly correlated with human perception. As a result, a meaningful evaluation of voice conversion systems requires the use of subjective measures to perceptually evaluate their conversion outputs.

2.3.1 Objective evaluation

Use of distance measures is most common for providing objective scores. One among them is spectral distortion (SD) which has been widely used to quantify spectral envelope conversions. For example, authors of [3] measured the ratio of spectral distortion

between the transformed and target speech and the source and target speech as follows:

$$R = \frac{SD(trans, tgt)}{SD(src, tgt)} \tag{2.9}$$

where R is the normalized distance $SD(trans, tgt)$ is the spectral distortion between the transformed and the target speaker utterance and $SD(src, tgt)$ is the spectral distortion between the source and the target speaker utterance.

A comparison of the performance of different types of conversion functions using a warped root mean square (RMS) log-spectral distortion measure was reported in [12]. Similar spectral distortion measures have been reported by other researchers [11] [53]. In addition, excitation spectrum, RMS-energy, F_0 and duration distances have also been used to measure excitation, energy, fundamental frequency and duration conversions [35].

Mel Cepstral Distortion (MCD) is another objective error measure used, which seems to have correlation with the subjective test results [54]. Thus MCD is used to measure the quality of voice transformation [13]. MCD is related to vocal characteristics and hence was an important measure to check the performance of mapping obtained by ANN/GMM network. MCD is essentially a weighted Euclidean distance defined as

$$MCD = (10/\ln 10) * \sqrt{2 * \sum_{i=1}^{25} (mc_i^t - mc_i^e)^2} \tag{2.10}$$

where mc_i^t and mc_i^e denote the target and the estimated Mel-cepstral coefficients, respectively.

2.3.2 Subjective evaluation

The objective of a voice conversion system is to transform an utterance of a speaker to sound as if spoken by the target speaker while maintaining the naturalness in speech.

Hence, in order to evaluate a VC system on these two scales, generally three types of subjective measures are used.

- ABX test
- Similarity test
- MOS score

- **ABX test:** In order to check if the converted speech is perceived as the target speaker, ABX tests are most commonly used where participants listen to source (A), target (B) and transformed (X) utterances and are asked to determine whether A or B is closer to X in terms of speaker identity. A score of 100% indicates that all listeners find the transformed speech closer to the target.

- **MOS score:** In addition to recognizability, the transformed speech is also generally evaluated in terms of naturalness and intelligibility by mean opinion score (MOS) tests. In this test, the participants are asked to rank the transformed speech in terms of its quality and/or intelligibility. This is similar to the similarity test, but the major difference lies in the fact that we concentrate on the speaker characteristics in similarity test and intelligibility in MOS score.

- **Similarity test:** The MOS score does not determine how similar the transformed speech and the target speech are. Hence, similarity measure is used, where the participants are asked to grade on a scale of 1 to 5 as to how close the transformed speech is to the target speaker's speech. A score of 5 means that the transformed and the target speech sound as if spoken by the same speaker and a scale of 1 indicates that both the utterances are from totally different speakers.

2.4 Applications

VC has mainly emerged as a technique to create new voices quickly for Text-To-Speech (TTS) systems. However it has a number of other interesting applications such as voice quality analysis, emotional speech synthesis and speech recognition. Fields such as speech-to-speech translation, education, health and entertainment have also developed applications using techniques involved in VC. Most commonly used TTS systems are based on unit selection. It is a technique which generates synthetic speech by selecting the most appropriate sequence of units from a database. The quality of a TTS system increases with the use of large databases [55] [56]. However with the use of VC techniques it is possible to build a TTS in a new voice with typically 30-50 utterances (10-15 minutes) from a new speaker. Hence, it is advantageous to employ VC for creating new TTS voice out of the existing ones [57], [58]. VC techniques have also emerged as a method for building multilingual TTS [59]. In this framework units from multiple languages are recorded by one speaker per language which are combined to improve the coverage of units. However TTS built on such a database has multiple speaker identities in the synthesized speech. Hence, a CLVC technique to transform the synthesized utterance to a particular target speaker is applied.

TTS system techniques are built to generate speech in different emotional modes such as excited, happy, sad or angry. The goal of these techniques is not only to generate speech in a given emotional state but also to have control of the amount of emotion to be generated. The techniques related to prosodic transformation are more appropriate to change the emotional state of a synthesized speech.

The problem of a speech recognition system is defined as the conversion of a spoken utterance into a textual sentence by machine. Such a system has to be sufficiently robust to allow use by a variety of speakers. using it. Hence VC could be used as a method for speaker normalization by converting all speaker's data onto a single speaker.

The motivation for building a CLVC framework is to be able to build a speech-to-

speech translation system [60] which involves transformation of speech spoken in one language to some other language. As the speaker may not know the target language, such a system is usually built to synthesize voice in some other speaker's voice whose utterances are recorded in the target language. VC techniques could be used in this case to transform the synthesized speech in target language to sound as if the source speaker is speaking it. Cross-lingual VC has also been applied to dubbing tasks in the film and music industries [61].

In the field of education, VC could be used to build a computer aided pronunciation training system. It is based on a study that for a second language learner it would be ideal if he has a feedback from a system which imitates his own voice but with a native speakers accent.

VC techniques could also be used to correct speech recorded by a speech-impaired/ disabled person leading to more natural and intelligible speech. Dysarthia [62] or laryngectomy [63] are examples of speech impairment.

Singing voice transformation and generating voices for virtual characters in a game are also some of the applications in which voice conversion techniques could be used. Authors of [64] were able to show that voice conversion techniques could be used in speaker de-identification, a case where we do not want to keep the individuality of the required speaker.

2.5 Summary

After a brief description of various methods of voice conversion and its applications, we understand that most of the methods aimed to find a better spectral transformation technique. We also observe that the GMM based methods are more often used and hence, we provide results of comparison of ANN and GMM based spectral transformation in Chapter 3. We also find that in order to address the issue of non-availability of parallel data, researchers have come up with methods that try to reduce the requirement of data

from the source speaker. However, referring to the state-of-the-art techniques proposed to address this issue, we infer that there is still a need for some data from the source speaker which motivates us to design a system which trains on only the target speaker data. Such a system will be described in Chapter 4.

CHAPTER 3

VOICE CONVERSION USING ARTIFICIAL NEURAL NETWORKS

From previous chapters, we understand that the very basic idea of voice conversion is to transform both the source and the excitation features. We have also observed that the focus in voice conversion is more on spectral transformation than the source feature transformation. From among various VC techniques, GMMs have been the most often used algorithm to build the transformation model. However, as we understand that the transformation of vocal tract features between two speakers is non-linear from the work reported in [15], we intend to use ANNs for voice conversion. We have exploited the mapping abilities of ANN to perform mapping of spectral features of a source speaker to that of a target speaker. A comparative study of voice conversion using ANN and the state-of-the-art Gaussian Mixture Model (GMM) is conducted. This chapter describes briefly the baseline framework for voice conversion and provides results comparing ANN and GMM based spectral transformation. The experiments in this chapter are done assuming an intra-lingual voice conversion framework with parallel training data i.e, the source speaker and the target speaker record a matching set of utterances.

3.1 Intra-lingual voice conversion

3.1.1 Database

Current voice conversion techniques need a parallel database [11] [13] [54] where the source and the target speakers record a matching set of utterances. The work presented here is carried out on CMU ARCTIC database consisting of utterances recorded by

seven speakers. Each speaker has recorded a set of 1132 phonetically balanced utterances [65]. The ARCTIC database includes utterances of SLT (US Female), CLB (US Female), BDL (US Male), RMS (US Male), JMK (Canadian Male), AWB (Scottish Male), KSP (Indian Male). It should be noted that about 30-50 parallel utterances are needed to build a voice conversion model [13]. Thus, for each speaker we took around 40 utterances (approximately 2 minutes) as training data and a separate set of 59 utterances (approximately 3 minutes) as testing data.

3.1.2 Feature extraction

To extract features from a speech signal, an excitation-filter model of speech is applied. Mel-cepstral coefficients (MCEPs) are extracted as filter parameters and fundamental frequency (F_0) estimates are derived as excitation features for every 5 ms [66]. The number of MCEPs extracted for every 5 ms is 25. Mean and standard deviation statistics of $\log(F_0)$ are calculated and recorded.

3.1.3 Alignment of parallel utterances

As the durations of the parallel utterances typically differ (as shown in Figure 3.1), dynamic time warping (or dynamic programming) is used to align MCEP vectors of the source and target speakers [12] [13]. Figure 3.1 is a plot of an utterance recorded by two speakers. The utterance consists of 18 phones, the boundaries of which are indicated by the vertical lines. It is very clear from this figure that the durations of phones in both the recorded utterances are different even though the spoken sentence is the same. Figure 3.2 shows that the durations of the two utterances can be matched after applying DTW.

After alignment, let x_t and y_t denote the source and target feature vectors at frame t respectively.

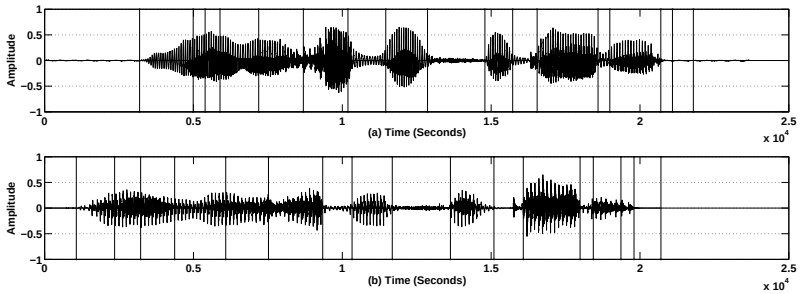

Figure 3.1: *Plot of an utterance recorded by two speakers showing that their durations differ even if the spoken sentence is the same. The spoken sentence is "Will we ever forget it" which has 18 phones "pau w ih l w iy eh v er f er g eh t ih t pau pau" according to the US English phoneset.*

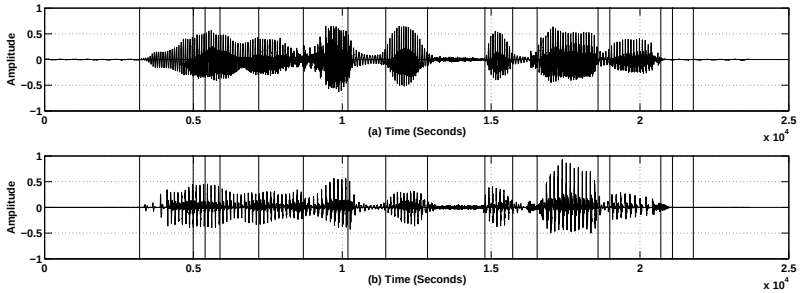

Figure 3.2: *Plot of an utterance recorded by two speakers showing that their durations match after applying DTW. The spoken sentence is "Will we ever forget it" which has 18 phones "pau w ih l w iy eh v er f er g eh t ih t pau pau" according to the US English phoneset.*

3.1.4 Process of training and testing/conversion

The training module of a voice conversion system to transform both the excitation and the filter parameters from a source speaker's acoustic space to a target speaker's acoustic space is as shown in Figure 3.3. Figure 3.4 shows the block diagram of various modules involved in a voice conversion testing process. In testing or conversion, the transformed MCEPs along with F_0 can be used as input to Mel Log Spectral Approximation (MLSA) [66] filter to synthesize the transformed utterance. For all the experiments done in this work, we have used pulse excitation for voiced sounds and random noise excitation for unvoiced sounds.

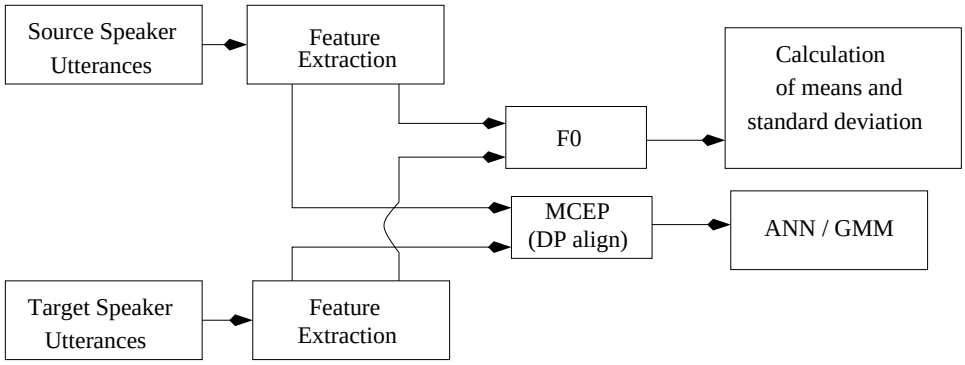

Figure 3.3: *Training module in voice conversion framework.*

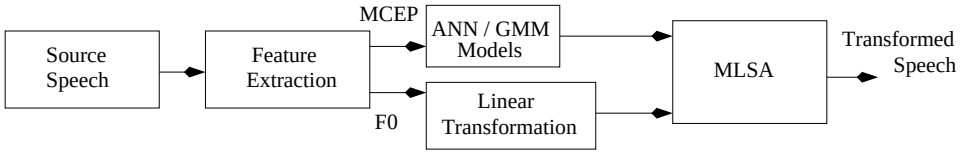

Figure 3.4: *Testing module in voice conversion framework.*

3.1.5 Spectral mapping using GMM

In GMM-based mapping [54] [67], the learning procedure aims to fit a GMM model to the augmented source and target feature vectors. Formally, a GMM allows the probability distribution of a random variable z to be modeled as the sum of M Gaussian components, also referred to as mixtures. Its probability density function $p(z_t)$ can be written as

$$p(z_t) = \sum_{m=1}^{M} \alpha_m \, \mathcal{N}(z_t; \mu_m^{(z)}, \Sigma_m^{(z)}) \quad \sum_{m=1}^{M} \alpha_m = 1, \; \alpha_m \geq 0 \tag{3.1}$$

where z_t is an augmented feature vector $[x_t^T \; y_t^T]^T$. The notation ^{T} denotes transposition of a vector. $\mathcal{N}(\mathbf{z}_t; \mu_m^{(z)}, \Sigma_m^{(z)})$ denotes the parameters of a Gaussian distribution and α_m denotes the prior probability with which the vector z_t belongs to the m^{th} component. $\Sigma_m^{(z)}$ denotes the covariance matrix and $\mu_m^{(z)}$ denotes the mean vector of the m^{th} component for the joint vectors. These parameters are represented as

$$\Sigma_m^{(z)} = \begin{bmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{bmatrix}, \; \mu_m^{(z)} = \begin{bmatrix} \mu_m^{(x)} \\ \mu_m^{(y)} \end{bmatrix}, \tag{3.2}$$

where $\mu_m^{(x)}$ and $\mu_m^{(y)}$ are the mean vectors of the m^{th} component for the source and the target feature vectors respectively. The matrices $\Sigma_m^{(xx)}$ and $\Sigma_m^{(yy)}$ are the covariance matrices, while $\Sigma_m^{(xy)}$ and $\Sigma_m^{(yx)}$ are the cross-covariance matrices, of the m^{th} component for the source and the target feature vectors respectively. The covariance matrices $\Sigma_m^{(xx)}$, $\Sigma_m^{(yy)}$, $\Sigma_m^{(xy)}$ and $\Sigma_m^{(yx)}$ are assumed to be diagonal in this thesis. The model parameters $(\alpha_m, \mu_m^{(z)}, \Sigma_m^{(z)})$ are estimated using Expectation Maximization (EM) algorithm.

The conversion process (also referred to as testing process) involves regression, i.e., given an input vector, x_t, we need to predict y_t using GMMs, which is calculated as shown in the equation below.

$$H(x_t) = E[y_t|x_t] = \sum_{m=1}^{M} h_m(x_t)[\mu_m^{(y)} + \Sigma_m^{(yx)}(\Sigma_m^{(xx)})^{-1}(x_t - \mu_m^{(x)})] \tag{3.3}$$

where

$$h_m(x_t) = \frac{\alpha_m \, \mathcal{N}(x_t; \mu_m^{(x)}, \Sigma_m^{(xx)})}{\sum_{n=1}^{M} \alpha_n \, \mathcal{N}(x_t; \mu_n^{(x)}, \Sigma_n^{(xx)})} \tag{3.4}$$

is the posterior probability that a given input vector x_t belongs to the m^{th} component.

In this work we have conducted GMM based VC experiments on the voice conversion setup built in FestVox distribution [68]. This voice conversion setup is based on the work done in [67], and supports the conversion considering 1) the feature correlation between frames (referred to as MLPG) and 2) the the Global Variance (GV) of spectral trajectory.

3.1.6 Spectral mapping using ANN

Artificial Neural Network (ANN) models consist of interconnected processing nodes, where each node represents the model of an artificial neuron, and the interconnection between two nodes has a weight associated with it. ANN models with different topologies perform different pattern recognition tasks. For example, a feed-forward neural

network can be designed to perform the task of pattern mapping, whereas a feedback network could be designed for the task of pattern association. A multi-layer feed forward neural network is used in this work to obtain the mapping function between the source and the target vectors.

Figure 3.5 shows the architecture of a four layer ANN used to capture the transformation function for mapping the acoustic features of a source speaker onto the acoustic space of a target speaker. The ANN is trained to map the MCEPs of a source speaker to the MCEPs of a target speaker, i.e., if $G(x_t)$ denotes the ANN mapping of x_t, then the error of mapping is given by $\varepsilon = \sum_t ||y_t - G(x_t)||^2$. $G(x_t)$ is defined as

$$G(x_t) = \widetilde{g}(w^{(3)}g(w^{(2)}g(w^{(1)}x_t))), \tag{3.5}$$

where

$$\widetilde{g}(\vartheta) = \vartheta, g(\vartheta) = a \ \tanh(b\,\vartheta). \tag{3.6}$$

Here $w^{(1)}, w^{(2)}, w^{(3)}$ represents the weight matrices of first, second and third hidden layers of ANN respectively. The values of the constants a and b used in tanh function are 1.7159 and $2/3$ respectively. A generalized back propagation learning [15] is used to adjust the weights of the neural network so as to minimize ε, i.e., the mean squared error between the desired and the actual output values. Selection of initial weights, architecture of ANN, learning rate, momentum and number of iterations are some of the optimization parameters in training an ANN [2]. Once the training is complete, we get a weight matrix that represents the mapping function between the spectral features of a pair of source and target speakers. Such a weight matrix can be used to transform a feature vector from the source speaker to a feature vector of the target speaker.

3.1.7 Mapping of excitation features

Our focus in this thesis is to get a better transformation of spectral features. Hence, we use the traditional approach of F_0 transformation as used in a GMM based transforma-

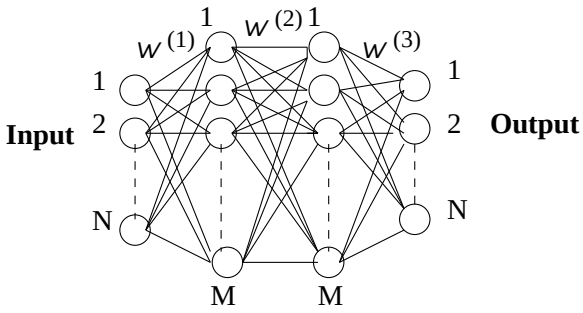

Figure 3.5: *An architecture of a four layered ANN with N input and output nodes and M nodes in the hidden layers.*

tion. A logarithm Gaussian normalized transformation [48] is used to transform the F_0 of a source speaker to the F_0 of a target speaker as indicated in the equation below

$$\log(F_{0\,conv}) = \mu_{tgt} + \frac{\sigma_{tgt}}{\sigma_{src}}(\log(F_{0\,src}) - \mu_{src}) \tag{3.7}$$

where μ_{src} and σ_{src} are the mean and variance of the fundamental frequency in logarithm domain for the source speaker, μ_{tgt} and σ_{tgt} are the mean and variance of the fundamental frequency in logarithm domain for the target speaker, $F_{0\,src}$ is the pitch of source speaker and $F_{0\,conv}$ is the converted pitch frequency.

3.1.8 Evaluation criteria for voice conversion

Subjective evaluation

Subjective evaluation is based on collecting human opinions as they are directly related to human perception, which is used to judge the quality of transformed speech. The popular tests are ABX test, MOS test and similarity test.

- *ABX Test:* For the ABX test, we present the listeners with a GMM transformed utterance and an ANN transformed utterance to be compared against X, which will always be a natural utterance of the target speaker. To ensure that a listener does not become biased, we shuffle the position of ANN/GMM transformed utterances i.e., A and B, with X always constant at the end. The listeners would be

asked to select either A or B, i.e., the one which they perceive to be closer to the target utterance.

- *MOS Test*: Mean Opinion Score (MOS) is another subjective evaluation where listeners evaluate the speech quality of the converted voices using a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad).

- *Similarity Test:* In similarity test, we present the listeners with a transformed utterance and a corresponding natural utterance of the target speaker. The listeners would be asked to provide a score indicating how similar the two utterances are in terms of speaker characteristics. The range of similarity test is also from 1 to 5 where a score of 5 indicates that both the recordings are from the same speaker and a score of 1 indicates that the two utterances are spoken by two different speakers.

Objective evaluation

Mel Cepstral Distortion (MCD) is an objective error measure known to have correlation with the subjective test results [54]. Thus MCD is used to measure the quality of voice transformation [13]. MCD is related to filter characteristics and hence is an important measure to check the performance of mapping obtained by an ANN/GMM model. MCD is computed as given in the equation below.

$$MCD = (10/\ln 10) * \sqrt{2 * \sum_{d=1}^{25} (mc_d^t - mc_d^e)^2} \tag{3.8}$$

where mc_d^t and mc_d^e denotes the d^{th} coefficient of the target and the transformed MCEP, respectively.

3.2 Experiments and results

3.2.1 Objective evaluation of a GMM based VC system

To build a GMM based VC system, we have considered two cases: 1) Transformation of SLT (US female) to BDL (US male) and 2) Transformation of BDL (US male) to SLT (US female). For both the experiments, the number of training utterances is 40 (approximately 2 minutes) and the testing is done on the test set of 59 utterances (approximately 3 minutes). The number of vectors for 40 training utterances in SLT and BDL are 23,679 and 21,820 respectively.

Table 3.1: Objective evaluation of GMM based VC system for various training parameters where Set 1: SLT to BDL transformation; Set 2: BDL to SLT transformation

No. of mixtures	**No. of params.**	**MCD [dB]**					
		Without MLPG		**With MLPG** (with GV)		**With MLPG** (Without GV)	
		Set 1	**Set 2**	**Set 1**	**Set 2**	**Set 1**	**Set 2**
32	6176	6.367	6.102	6.547	6.072	6.152	5.823
64	12352	6.336	6.107	6.442	6.015	6.057	5.762
128	24704	6.348	6.068	6.389	5.907	6.017	5.682

Table 3.1 provides the MCD scores computed for SLT-to-BDL and BDL-to-SLT respectively for increasing number of Gaussians. It could be observed that the MCD scores decrease with the increase in the number of Gaussians, however, it should be noted that the increase in the number of Gaussians also increases the number of parameters in the GMM. The number of parameters for a GMM based system with diagonal covariance matrix is computed as follows:(((dimension of mean vector + dimension of variance vector)*No. of Gaussians) + No. of Gaussians). With the use of diagonal covariance matrix, the number of parameters in the GMM with 64 and 128 Gaussian components is 12,352 and 24,704 respectively. We can also observe that the GMM based conversion with MLPG performs better than that of the GMM based system

without MLPG. However, the GMM based VC system with MLPG and without GV produced lesser MCD scores than the GMM based VC system with MLPG and with GV. While GV seemed to improve the quality of transformed speech based on human listening tests, it is not clear from [67] whether it also improves the score according to MCD computation. Considering the number of parameters used in GMM model, we have used the GMM based VC system with 64 Gaussian components (with MLPG and without GV) for further comparison with an ANN based VC system.

3.2.2 Objective evaluation of an ANN based VC system

To build an ANN based VC system, we have considered two cases 1) SLT-to-BDL and 2) BDL-to-SLT. For both the experiments, the number of training utterances is 40 (approximately 2 minutes) and the testing is done on the test set of 59 utterances (approximately 3 minutes).

Table 3.2 provide MCD scores for SLT-to-BDL and BDL-to-SLT respectively for different architectures of ANN. In this work, we have experimented with 3-layer, 4-layer and 5-layer ANNs. The architectures are provided with the number of nodes in each layer and the activation function used for that layer. For example, 25L 75N 25L means that it is a 3-layer network with 25 input and output nodes and with 75 nodes in the hidden layer. Here, L represents ”linear” activation function and N represents ”tangential (tanh$(.)$)” activation function. Given an ANN architecture, the No. of parameters to be computed is calculated as follows: Suppose the ANN architecture is 25L 50N 50N 25L, the number of parameters is (25*50)+(50*50)+(50*25)+(50+50+25) = 5125. From Table 3.2, we see that the four layered architecture 25L 50N 50N 25L (with 5125 parameters) provides better results when compared with other architectures. Hence, for all the remaining experiments reported in this chapter, a four layer architecture is used.

In order to determine the effect of number of parallel utterances used for training the voice conversion models, we performed experiments by varying the training data from

Table 3.2: MCD obtained on the test set for different architectures of an ANN model. (No. of iterations: 200, Learning Rate: 0.01, Momentum: 0.3) Set 1: SLT to BDL; Set 2: BDL to SLT

S.No	ANN architecture	No. of params.	MCD [dB]	
			Set 1	Set 2
1	25L 75N 25L	3850	6.147	5.652
2	**25L 50N 50N 25L**	**5125**	**6.048**	**5.504**
3	25L 75N 75N 25L	9550	6.147	5.571
4	25L 75N 4L 75N 25L	4529	6.238	5.658
5	25L 75N 10L 75N 25L	5435	6.154	5.527
6	25L 75N 20L 75N 25L	6945	6.151	5.517

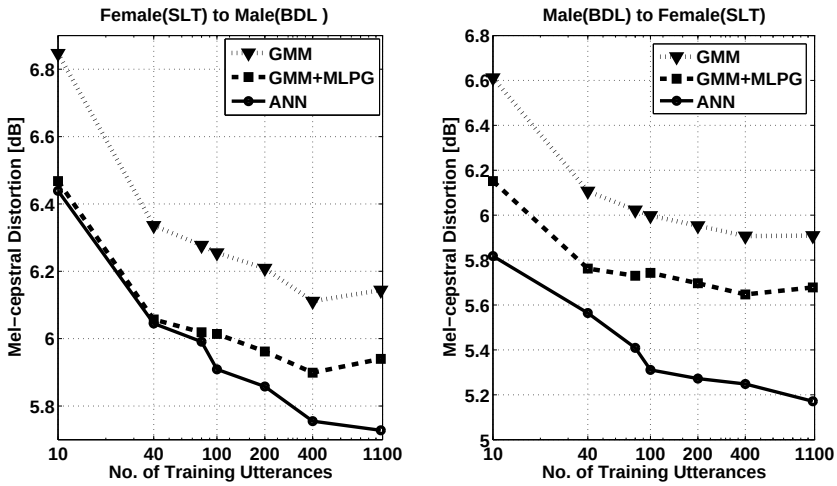

Figure 3.6: *MCD scores for ANN, GMM+MLPG and GMM (without MLPG) based VC systems computed as a function of number of utterances used for training. The results for GMM based VC systems are obtained using 64 mixture components.*

10 to 1073 parallel utterances. Please note that the number of test utterances was always 59. Figure 3.6 shows the MCD scores for ANN, GMM + MLPG and GMM (without MLPG) based VC systems computed as a function of number of utterances used for training. From Figure 3.6, we could observe that as the number of training utterances increase, the MCD values obtained by both GMM and ANN models decrease.

3.2.3 Subjective evaluation of GMM and ANN based VC systems

In this section we provide subjective evaluations for ANN and GMM based voice conversion systems. For these tests, we have made use of voice conversion models built from 40 parallel utterances, as it was shown that this modest set produces good enough transformation quality in terms of objective measure. We conducted MOS, ABX and similarity tests to evaluate the performance of the ANN based transformation against the GMM based transformation. It has to be noted that all experiments with GMM use static and delta features but the experiments with ANN use only the static features.

A total of 32 subjects were asked to participate in the four experiments listed below. Each subject was asked to listen to 10 utterances corresponding to one of the experiments. Figure 3.7(a) provides the MOS scores for 1) ANN, 2) GMM + MLPG and 3) GMM (without MLPG) based VC systems. Figure 3.7(b) provides the results of ABX test for the following cases:

4) BDL to SLT using ANN + (GMM + MLPG)

5) SLT to BDL using ANN + (GMM + MLPG)

6) BDL to SLT using ANN + GMM

7) SLT to BDL using ANN + GMM

The MOS scores and ABX tests indicate that the ANN based VC system performs as good as that of the GMM based VC system. The MOS scores also indicate that the transformed output from the GMM based VC system with MLPG was perceived to be better than that of the GMM based VC system without MLPG.

A similarity test is also performed between the output of the ANN/GMM based VC system and the respective natural utterances of the target speaker. The results of this similarity test are provided in Table 3.3, which indicate that the ANN based VC system seems to perform better or as good as that of the GMM based VC system.

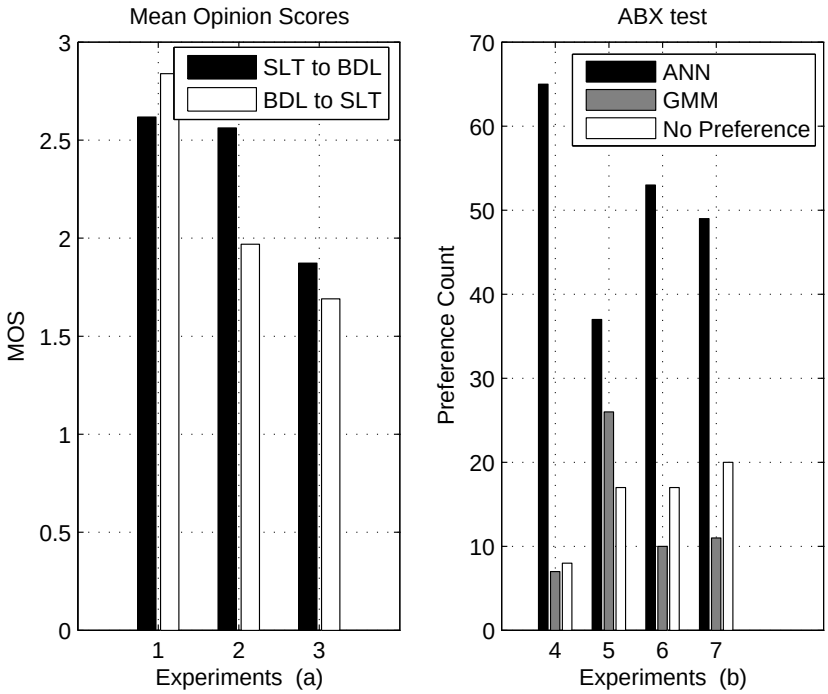

Figure 3.7: ***(a)*** *- MOS scores for 1: ANN, 2: GMM+MLPG, 3: GMM.* ***(b)*** *ABX results for 4: ANN, GMM+MLPG(M->F), 5: ANN, GMM+MLPG(F->M), 6: ANN, GMM(M->F), 7: ANN, GMM(F->M)*

The significance of difference between the ANN and the GMM+MLPG based VC systems for MOS and similarity scores was tested using hypothesis testing based on Student t-test, and the level of confidence indicating the difference was found to be greater than 95%.

Table 3.3: Average similarity scores between transformed utterances and the natural utterances of the target speaker.

Transformation Method	**Avg. Similarity Score**	
	SLT to BDL	BDL to SLT
ANN	2.93	3.02
GMM + MLPG	1.99	2.56

3.2.4 Experiment on multiple speaker pairs

In order to show that the ANN based transformation can be generalized over different databases, we have provided MOS and MCD scores for voice conversion performed

for 10 different pairs of speakers as shown in Figure 3.8. While MCD values were obtained over the test set of 59 utterances, the MOS scores were obtained from 16 subjects, each performing the listening tests on 10 utterances. An analysis drawn from these results show that inter-gender voice transformation (ex: male to female) has an average MCD and a MOS score of 5.79 and 3.06 respectively while the intra-gender (ex: male to male) voice transformation has an average MCD and a MOS score of 5.86 and 3.0 respectively. Another result drawn from the above experiments indicates that the transformation performance between two speakers with the same accent is better than that when compared with performance on speakers with different accent. For example, the voice transformation from SLT (US accent) to BDL (US accent) obtained an MCD value of 5.59 and a MOS score of 3.17, while the voice transformation from BDL (US accent) to AWB (Scottish accent) obtained an MCD value of 6.04 and a MOS score of 2.8.

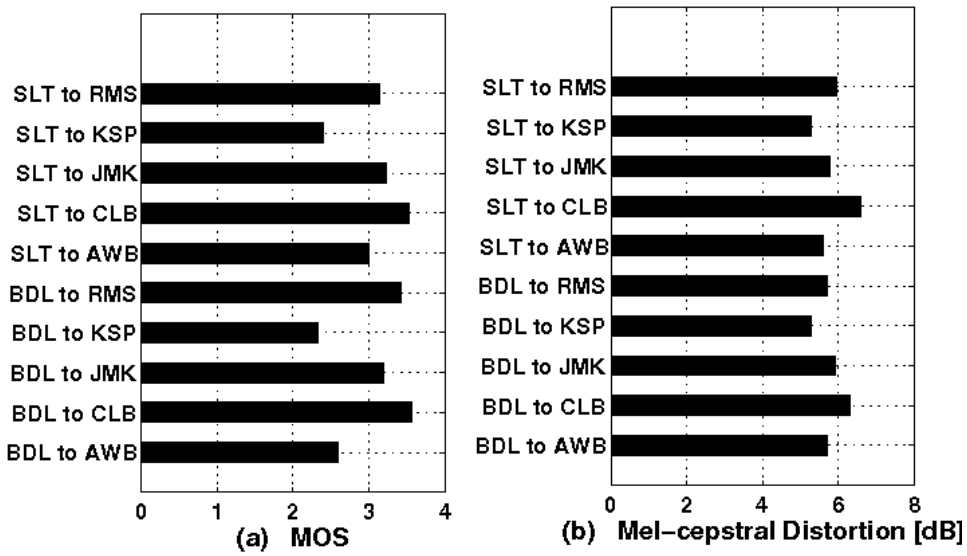

Figure 3.8: ***(a)*** *MOS and* ***(b)*** *MCD scores for ANN based VC system on 10 different pairs of speakers*

3.3 Enhancements to voice conversion using ANN

In order to enhance the performance of spectral mapping by ANNs, we investigated two different methods. All the experiments in this section are designed based on the use of parallel training data. The results of these experiments are provided on the test set of 59 utterances.

3.3.1 Appending deltas

The GMM based approach explained in Section 3.1.5 appends dynamic features to the static features [54] [21] [20]. In section 3.2, we have compared the GMM based system with deltas with the ANN based system without deltas and hence we wanted to find out whether the use of deltas would further improve the performance of an ANN based system.

In this context we performed an experiment on SLT (female) to BDL (male) transformation where the model is trained with deltas and on varying number of parallel training utterances. A set of three experiments were conducted and the architectures of ANN used in these experiments are as follows:

1. 25L 50N 50N 25L: static features.
2. 50L 100N 100N 50L: static and delta features.
3. 75L 150N 150N 75L: static, delta and acceleration/delta-delta features.

The MCD scores obtained for these three experiments are provided in Table 3.4. It can be observed that the ANN transformation with deltas is better than the ANN based transformation without deltas. The results using delta-delta features are also provided in Table 3.4. It could be observed that use of delta-delta features further reduces the MCD score for ANN based spectral mapping. The set of 40 training utterances used in this experiment is different than the one used in Section 3.2.2 and hence we find minor differences in the MCD scores for static features.

Table 3.4: Results of appending deltas and delta-deltas of MCEPs for (SLT(female) to BDL(male) transformation)

No. of training utterances	static features MCD [dB]	deltas MCD [dB]	delta-delta MCD [dB]
40	6.118	6.117	6.088
100	6.018	5.995	5.905
200	5.858	5.854	5.836
400	5.755	5.750	5.695

3.3.2 A hybrid model

Even though the ANNs perform better spectral transformation than that of GMMs, it would be interesting to find out if combining these approaches to build a hybrid conversion model could improve the transformation performance. Hence, we came up with a hybrid technique which combines the output of ANN and GMM techniques to predict a new set of transformed vectors.

There are two phases in this hybrid model. In the first phase ANN and GMM are trained as explained in Section 3. After the models are built, the transformed output vectors are obtained for the training files. Let $\mathbf{y_i^a}$ and $\mathbf{y_i^g}$ denote the transformed vectors for input $\mathbf{x_i}$ using ANN and GMM respectively. We experimented the use of ANNs with context size of 3 as they performed best among all the conducted experiments, we used GMM with deltas, MLPG, No GV and with 64 mixture components. Each of these modules of phase 1 will predict output vector with 25 dimensions each. In second phase these transformed vectors are appended to form a 50 dimensional input vector $[\mathbf{y_i^a}, \mathbf{y_i^g}]$ which is given to train another neural network to map these 50 dimensional vectors to its original target speaker 25 dimensional MCEPs $\mathbf{x_i}$. However, contextual features with context size of 3 was also used and hence the mapping was between features of dimension 350 and 175. The ANN architecture chosen for this mapping was 350L 700N 700N 175L. The output is again scaled down to 25 dimensions before comparing with the actual MCEP values for MCD computation.

The idea is that the second neural network model could perform a weighted combi-

nation of $[\mathbf{y_i^a}, \mathbf{y_i^g}]$ to predict the transformed vector as close as possible to $\mathbf{x_i}$. Figure 3.9 shows the way in which a hybrid model of ANN-GMM is built.

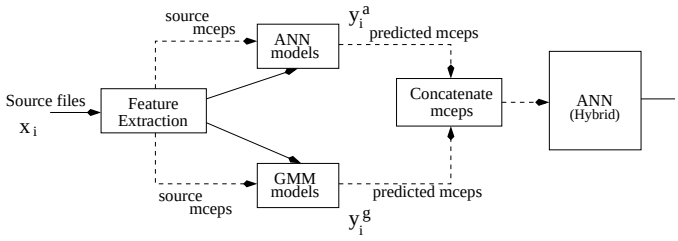

Figure 3.9: *Block diagram for the hybrid approach*

This experiment was trained on 40 utterances for transformation between SLT and BDL transformation. Experiments were conducted with an increasing number of input speech utterances. Figure 3.10 shows that the output from the ANN-GMM hybrid approach is better than using only the ANN based approach. It should be noted that the training data set used in this experiment is different from that used in Section 3, and hence MCDs values obtained in Section 3 and in this experiment differ for ANN and GMM based systems. From the experiments done for enhancing the spectral mapping using ANN, we could observe that use of enhancement methods provide better performance than the base-line ANN based spectral mapping, although they add a little extra computation time.

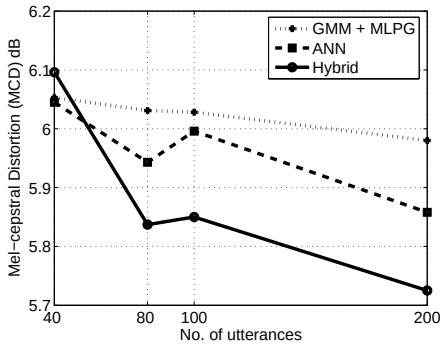

Figure 3.10: *MCD scores for the hybrid approach with increasing training data*

3.3.3 Transformation with use of contextual features

The use of deltas and delta-delta coefficients are computed over a context of 3 frames, and provide slope and acceleration coefficients of MCEPs [69]. Instead of computing slope and acceleration coefficients, we wanted to investigate the effect of augmented feature vectors, i.e., append MCEPs from previous and next frames to the MCEPs of current frame, and provide these augmented features as input to train the ANN model.

In this context, we performed an experiment on SLT (female) to BDL (male) transformation, where the model is trained on varying context size and varying number of parallel training utterances. A context size of one indicates that MCEPs from one left and one right frame are appended to MCEPs of the current frame. The results of SLT to BDL transformation are provided in Figure 3.11, where a plot showing the MCD score with increasing number of training utterances and context size is provided.

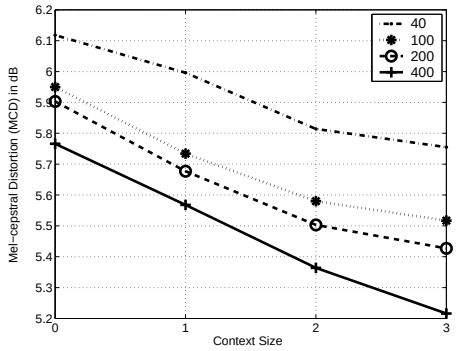

Figure 3.11: *Graph of MCD as a function of context size for varying number of training utterances for SLT (female) to BDL (male). Context 0 indicates the baseline performance.*

Figure 3.11 shows that the MCD score decreases with the increase in context size from 0 to 3 (i.e., 3 left and 3 right frames). The MCD score at the context size of 0 in Figure 3.11 indicates the base line performance as explained in Section 3.2.2. The ANN architectures used for context size of 1, 2 and 3 are 75L 225N 225N 75L, 125L 375N 375N 125L and 175L 525N 525N 175L, respectively. From Figure 3.11, it could also be observed that increase in the number of training utterances from 40 to 200 leads to a decrease in MCD scores and thus improves the performance of the ANN based spectral mapping.

From the experiments conducted in Section 3.3.1 and 3.3.3, we could observe that the use of deltas, acceleration coefficients and contextual features improves the performance of an ANN based VC system. However, an increase in the dimensionality of feature vectors also increases the size of an ANN architecture and the computational load in training and testing of a VC system.

3.4 Summary

In this chapter, we have exploited the mapping abilities of ANN and it was shown that ANN can be used for spectral transformation in the voice conversion framework on a continuous speech signal. The usefulness of ANN has been demonstrated on different pairs of speakers. Comparison between ANN and GMM based transformation has shown that the ANN based spectral transformation yields better or as good results as GMM based system in both the objective and subjective evaluations. Hence, we conclude that no smoothing techniques are required when using ANNs for spectral transformation. We also discuss various possibilities of improving the spectral transformation performance given that the ANN based spectral transformation is our baseline system. Appending deltas to obtain better performance has been used in various experiments and is well known to improve the performance of a system when compared with a system built on only the static features. The results obtained in our tests also conform to this result. We propose the use of contextual features and a hybrid ANN-GMM approach whose results are also found to be much better when compared to the use of only ANNs for spectral transformation.

CHAPTER 4

FRAMEWORK FOR SOURCE-SPEAKER INDEPENDENT VOICE CONVERSION

The methods described so far rely on availability of parallel data for training the models. Hence each time a new speaker wants to transform his/her voice to a target speaker, he/she will have to record all the utterances of the target speaker so that a parallel data can be obtained. This is a costly and time consuming task. Hence, there is a need to come up with an approach which will avoid the source speaker having to give any speech samples for training the voice conversion model. There are a few approaches which have tried to reduce the contribution from the source speaker. These methods include using a speech recognizer [23], a unit selection algorithm [19] class mapping [24], creating pseudo parallel corpus [10] and using adaptation techniques [25] as explained in Chapter 2. However, all these methods still need both the speakers data (though not parallel utterances). Hence, there is a need to design an algorithm that will capture speaker specific characteristics and hence use only the target speaker data. Such an algorithm which needs only target speaker data can be used in both ILVC and CLVC frameworks. In this chapter, we propose a voice conversion method using ANN which captures speaker-specific characteristics of a target speaker. Such a model avoids the need for speech data from a source speaker and hence could be used to transform arbitrary speaker including a cross-lingual speaker.

4.1 Many-to-one mapping

We propose an approach where multiple speakers acoustic space is mapped onto a single target speaker acoustic space during training and the obtained model is used to

transform any arbitrary speakers speech. A similar approach was followed in [21], where speaker adaptation technique was also used. However, we have not performed any speaker adaptation techniques in this work.

Our goal is to come up with an approach where the need for a source speaker's training data is totally eliminated. ARCTIC database is a collection of multiple speakers uttering the same sentences. The experiments in this section are undertaken with an assumption that the source speaker is SLT and the target speaker is BDL. To start with, we train an ANN model with the input and output training data as BDL MCEPs from 40 utterances. However, during testing, we check if SLT can be transformed to BDL. The results are provided in Table 4.1.

Finally we train an ANN model to map 40 utterances from each of BDL, AWB, KSP, CLB, RMS, JMK to the target speaker BDL. This kind of conversion should work reasonably well as it models transformation of multiple speakers to a particular target speaker and hence capture the acoustic variations. During testing, SLT utterances are given to this model to convert them onto BDL acoustic space. We also experimented with training an ANN model to transform BDL MCEPs to BDL MCEPs with the aim of capturing target speaker-specific characteristics. However, we observed that an ANN model trained to map multiple source speaker to a single target speaker performs better than a self mapping network. Results of both these experiments are provided in Table 4.1.

Table 4.1: Many to one mapping

Trained model	**No. of training utterances**	**Testing input speaker**	**MCD [dB]**
BDL to BDL	40	SLT	8.763
(BDL, RMS, AWB, KSP, CLB, JMK) to BDL	(6 speakers * 40 each) = 240	SLT	6.674

From Table 4.1 we see that the MCD scores are better when the models are trained to map multiple source speakers onto a single target speaker. Informal listening tests

showed the transformed voice possesses the voice characteristics of the target speaker and is intelligible.

4.2 Models capturing speaker-specific characteristics

So far we have discussed VC approaches which rely on existence of parallel data from the source and the target speakers. There have been approaches proposed in [24], [44], [19], [25], which avoid the need for parallel data, however still require speech data (though non-parallel) from source speakers *apriori* to build a voice conversion model. Such approaches cannot be applied in situations, where an arbitrary user intends to have his/her voice transformed to a pre-defined target speaker, without recording anything *apriori*. In this section, we propose a voice conversion approach using an ANN model which captures speaker-specific characteristics of a target speaker. Such an approach does not require speech data from a source speaker and hence could be used to transform an arbitrary speaker including a cross-lingual speaker. The idea behind capturing the speaker-specific characteristics using an ANN model is as follows. Let l_q and s_q be two different representations of a speech signal from a target speaker q. A mapping function $\Omega(l_q)$ could be built to transform l_q to s_q. Such a function would be specific to the speaker q and could be considered as capturing the essential speaker-specific characteristics. The choice of representation of l_q and s_q plays an important role in building such mapping networks and in their interpretation. If we assume that l_q represents linguistic information, and s_q represents linguistic and speaker information, then a mapping function from l_q to s_q should capture speaker-specific information in the process. The interpretation of order of Linear Prediction (LP) could be applied in deriving l_q and s_q. A lower order (≤ 6) LP spectrum captures first few formants and mostly characterizes the message (or linguistic) part of the signal, while a higher order (≥ 12) LP spectrum captures more details in the spectrum and hence captures message and speaker characteristics [70]. Thus l_q being represented by a lower order LP spectrum of first few formants could be interpreted as speaker independent representation

of the speech signal, and s_q represented by the MCEPs could be interpreted as carrying message and speaker information. An ANN model could be trained to minimize the error $||s_q^{'} - s_q||$, where $s_q^{'} = \Omega(l_q)$.

In this work, l_q is represented by six formants, their bandwidths and delta features. The formants, bandwidths, F_0 and probability of voicing are extracted using the ESPS toolkit [71] . The formants also undergo a normalization technique such as vocal tract length normalization explained in Section 4.2.1. s_q is represented by traditional MCEP features as it would allow us to synthesize using MLSA synthesis technique. The MLSA synthesis technique generates a speech waveform from the transformed MCEPs and F0 values using pulse excitation or random noise excitation [66]. An ANN model is trained to map l_q to s_q using backpropagation learning algorithm. Once the model is trained, it could be used to convert l_r to $s_q^{'}$ where l_r could be from any arbitrary speaker r.

4.2.1 Vocal tract length normalization

VTLN is a speaker normalization technique that tries to compensate for the effect of speaker-dependent vocal tract lengths by warping the frequency axis of the magnitude spectrum. Apart from the use of VTLN in speech recognition, VTLN has also been used in voice conversion [24] [44] [19].

Following the work in [72], we estimate the warp factors using pitch information and modify both formants and bandwidths. A piece-wise linear warping function as described in the equation 4.1 is used.

$$f_t' = \begin{cases} k f_t & : \quad f_t \leq F_{0\,t} \\ kF_{0\,t} - \frac{f_N - kF_{0\,t}}{f_N - F_{0\,t}} & : \quad F_{0\,t} < f_t < f_N \end{cases} \tag{4.1}$$

where $k = 1 - 0.002(F_{0\,t} - f_{mean})$, f_t is the formant / bandwidth frequency of frame t to be normalized , $F_{0\,t}$ is the pitch value of the frame t and f_N is the sampling frequency. f_{mean} is the mean pitch of the target speaker.

Figures 4.1 and 4.2 show the LPC spectrum of a phone /aa/ spoken by two male and female speakers respectively. The plots on the left columns indicate the original spectrum and the plots on the right columns indicate the transformation after VTLN. It is clear from these figures that the spectrum is normalized to a similar domain and hence our objective is achieved.

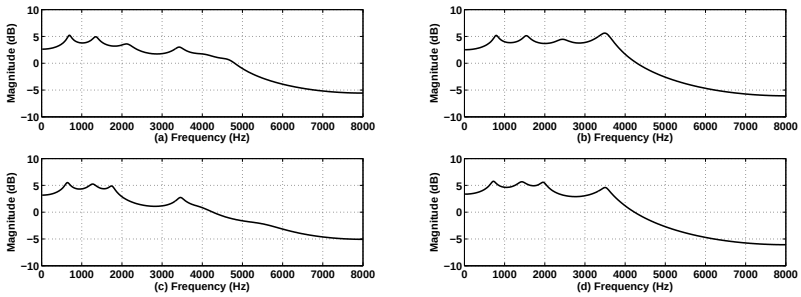

Figure 4.1: *Plot of LPC spectrum of two male speakers with the original spectrum on the left column and normalized spectrum on the right column.*

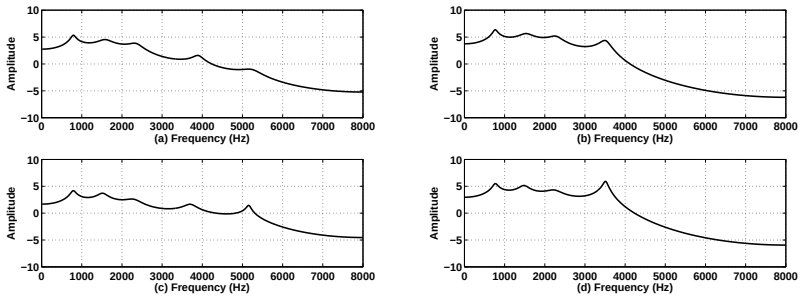

Figure 4.2: *Plot of LPC spectrum of two female speakers with the original spectrum on the left column and normalized spectrum on the right column.*

4.2.2 Error correction network

We introduce a concept of error correction network which is essentially an additional ANN network, used to map the predicted MCEPs to the target MCEPs so that the final output obtained features represent the target speaker in a better way. The block diagram for error correction network is shown in Figure 4.3. Once $s_q^{'}$ are obtained, they are given as input to second ANN and it is trained to reduced the error $||s_q^{'} - s_q||$. Such a network acts as error correction network to correct any errors made by first ANN. Let $s_q^{''}$ denote the output from error correction network. It is observed that while the

MCD values of $s_q^{'}$ and $s_q^{''}$ do not vary much, the speech synthesized from $s_q^{''}$ was found to be smoother than that of speech synthesized from $s_q^{'}$. To train the error correction network, we use 2-D features i.e., feature vectors from 3 left frames, and 3 right frames are added as context to the current frame. Thus the ANN model is trained with 175 dimensional vector (25 dimension MCEPs * (3+1+3)). The architecture of this error correction network is 175L 525N 525N 175L.

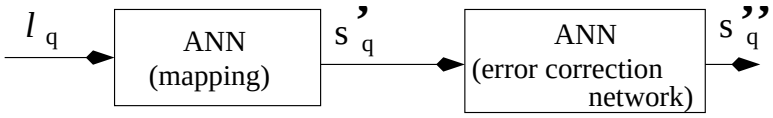

Figure 4.3: *A block diagram of an error correction network*

4.2.3 Experiments with parallel data

As an initial experiment, we used parallel data of BDL and SLT. Features representing l_r were extracted from BDL speaker and were mapped onto s_q of SLT. This experimentation was done mainly to obtain a benchmark performance for the experiments which map l_q to s_q (as explained in Section 3.2.2).

The features representing l undergo a VTLN (as discussed in Section 4.2.1), to normalize the speaker effect on the message (or linguistic) part of the signal. However, in this experiment, the mapping is done between BDL's l_r to SLT's s_q. The process of training such a voice conversion model is similar to the process explained in Section 3.2. Hence, VTLN was not performed on the features representing l_r in this experiment.

Table 4.2: Results of source speaker (SLT-female) to target speaker (BDL-male) transformation with training on 40 utterances of source formants to target MCEPs on a parallel database. Here **F** represents Formants, **B** represents Bandwidths, Δ and $\Delta\Delta$ represents delta and delta-delta features computed on **ESPS** features respectively. **UVN** represents unit variance normalization.

S.No	Features	ANN architecture	MCD [dB]
1	4 F	4L 50N 12L 50N 25L	9.786
2	4 F + 4 B	8L 16N 4L 16N 25L	9.557
3	4 F + 4 B + UVN	8L 16N 4L 16N 25L	6.639
4	4 F + 4 B + Δ + $\Delta\Delta$ + UVN	24L 50N 50N 25L	6.352
5	F_0 + 4 F + 4 B + UVN	9L 18N 3L 18N 25L	6.713
6	F_0 + 4 F + 4 B + Δ + $\Delta\Delta$ + UVN	27L 50N 50N 25L	6.375
7	F_0 + Prob. of Voicing + 4 F + 4 B + Δ + $\Delta\Delta$ + UVN	30L 50N 50N 25L	6.105
8	**F_0 + Prob. of voicing + 6 F + 6 B + Δ + $\Delta\Delta$ + UVN**	**42L 75N 75N 25L**	**5.992**
9	**(F_0 + Prob. of voicing + 6 F + 6 B + Δ + $\Delta\Delta$ + UVN) + (3L3R MCEP to MCEP error correction)**	**(42L 75N 75N 25L) + (175L 525N 525N 175L)**	**5.615**

Training was done to map BDL formants to SLT MCEPs with only 40 utterances. Testing was done on a set of 59 utterances. Table 4.2 shows the different representations of l_r and their effect on MCD values. These different representations include combination of different number of formants and their bandwidths, delta and acceleration coefficients of formants and bandwidths, pitch and probability of voicing. From the results provided in Table 4.2 we can observe that experiment 9 (which uses six formants, six bandwidths, probability of voicing, pitch along with their delta and acceleration coefficients) employing an error correction network provided better results in terms of MCD values. These results are comparable with the results of voice conversion with BDL MCEPs to SLT MCEPs mapping as found in Section 3.2.2.

4.2.4 Experiments using target speaker's data

In this work, we built an ANN model which maps l_q features of SLT onto s_q features of SLT. Here l_q extracted from SLT utterances is represented by six formants, six bandwidths, F_0, probability of voicing and their delta and acceleration coefficients as shown in feature set for experiment 9 in Table 4.2. The formants and bandwidths representing

l_q undergo VTLN to normalize speaker specific characteristics. s_q is represented by MCEPs extracted from SLT utterances. We use the concept of error correction network to improve the smoothness of the converted voice.

Figure 4.4 provides the results for mapping l_r (where r = BDL, RMS, CLB, JMK voices) onto the acoustic space of SLT. To perform this mapping the voice conversion model is built to map l_q to s_q (where q = SLT) is used. To perform VTLN, we have used the mean pitch value of SLT. Hence all the formants of source speaker are normalized with VTLN using mean of SLT F_0 and then are given to ANN to predict the 25 dimensional MCEPS. Similar results where the conversion model is built by capturing BDL speaker-specific features are also provided in Figure 4.4.

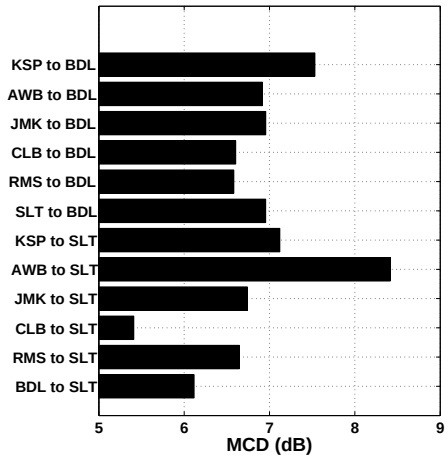

Figure 4.4: *A plot of MCD scores obtained between multiple speaker pairs with SLT or BDL as target speakers. The models are built from a training data of 24 minutes and tested on 59 utterances (approximately 3 min).*

We also performed listening tests whose results are provided in Table 4.3 for MOS scores and similarity tests. For the listening tests we chose 3 utterances randomly from each of the transformation pairs. Table 4.3 provides a combined output of all speakers transformed to target speaker (SLT or BDL). There were 10 listeners who participated in the evaluations tests. The MOS scores and similarity test results are averaged over 10 listeners.

The results shown in Figure 4.4 and Table 4.3 indicate that voice conversion models built by capturing speaker-specific characteristics using ANN models are useful. As

Table 4.3: Subjective evaluation of voice conversion models built by capturing speaker-specific characteristics

Target Speaker	MOS	Similarity tests
BDL	2.926	2.715
SLT	2.731	2.47

this approach do not need any utterances from source speaker to train a voice conversion model we can use this type of model to perform cross-lingual voice conversion. Figure 4.5 shows the effect of amount of training data in building the ANN models capturing speaker-specific characteristics. It could be observed that the MCD scores tend to decrease with the increase in the amount of training data.

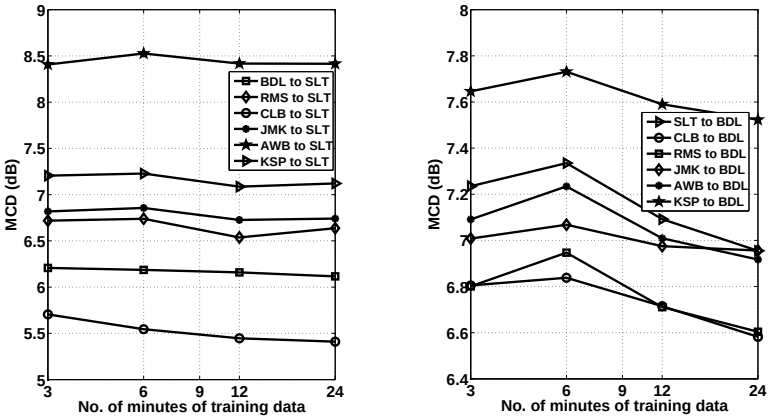

Figure 4.5: *A plot of MCD v/s Data size for different speaker pairs and with SLT or BDL as the target speaker.*

4.2.5 Experiments on multiple speakers database

To test the validity of the proposed method, we conducted experiments on other databases from the ARCTIC set, such as RMS, CLB, JMK, AWB and KSP. The training for all these experiments was conducted on 6 minutes of speech data form each of the database. However, the testing was done on the standard set of 59 utterances. The MCD scores provided in Table 4.4 indicate that the methodology of training an ANN model to capture speaker-specific characteristics for voice conversion could be generalized over different datasets.

Table 4.4: Performance of voice conversion model built by capturing speaker-specific features are provided with MCD scores. Entries in the first column represent source speakers and the entries in the first row represent target speakers. All the experiments are trained on 6 minutes of speech and tested on 59 utterances or approximately 3 minutes of speech.

Source \ Target	RMS	CLB	AWB	KSP
BDL	6.260	6.137	6.558	6.820
SLT	7.430	5.791	6.354	7.278
CLB	7.066	NA	6.297	7.166
JMK	6.617	6.616	6.224	6.878
RMS	NA	6.716	6.251	6.891
AWB	6.847	6.517	NA	6.769
KSP	7.392	7.239	6.517	NA

4.2.6 Application to cross-lingual voice conversion

Table 4.5: Subjective results of cross-lingual transformation done using conversion model built by capturing speaker-specific characteristics. 10 utterances from each of Telugu (NK), Hindi (PRA) and Kannada (LV) speakers are transformed into BDL male speaker's voice

Source Speaker	**Target Speaker**	**MOS**	**Similarity tests**
NK (Telugu)	BDL (English)	2.88	2.77
PRA (Hindi)	BDL (English)	2.62	2.15
LV (Kannada)	BDL English	2.77	2.22

Cross-lingual voice conversion is a task where the language of the source and the target speakers is different. In the case of speech-to-speech translation system, a source speaker may not know the target language. Hence, to convey information in his/her voice in the target language, cross-lingual voice conversion assumes importance. The availability of parallel data is difficult for cross-lingual voice conversion. One solution is to perform a unit selection approach [24] [44] [19] to find units in target speaker utterances that are close to the source speaker or use utterances recorded by a bi-lingual speaker [43]. Our solution to cross-lingual voice conversion is to employ the ANN model which captures speaker-specific characteristics.

In this context, we performed an experiment to transform three female speakers (NK, PRA, LV) speaking Telugu, Hindi and Kannada respectively into a male voice speaking English (US male - BDL). Our goal here is transform NK, PRA and LV voices to BDL voice and hence the output will be as if BDL were speaking in Telugu, Hindi and Kannada respectively. We make use of BDL models built in Section 4.2.4 to capture speaker-specific characteristics. Ten utterances from NK, PRA, LV voices were transformed into BDL voice and we performed MOS test and similarity test to evaluate the performance of this transformation. Table 4.5 provides the MOS and similarity test results averaged over all listeners. There were 10 native listeners of Telugu, Hindi and Kannada who participated in the evaluations tests. The MOS scores in Table 4.5 indicate the transformed voice was intelligible. The similarity tests indicate that cross-lingual transformation could be achieved using ANN models, and the output is intelligible and possess the characteristics of BDL voice.

4.3 Summary

The focus in this chapter was to design a framework which does not use any of the source speaker recordings. Use of a many-to-one approach does transform the source speaker speech to the target speaker, however, its performance depends on the availability of multiple pre-stored speakers. Our second approach which does not need any data other than the target speakers data is a novel approach which also seems to perform well in the case of both intra-lingual and cross-lingual voice conversion.

CHAPTER 5

Conclusion and future work

5.1 Conclusion

Most of the current voice conversion systems need training data from a source speaker and a target speaker. Such training data could be either parallel, where both the speakers record the same set of utterances or non-parallel, where the utterances recorded by both the speakers are different. It is known that the use of parallel training data provides better results when compared to the systems using non-parallel training data. Machine learning techniques such as VQ, GMM, HMM, ANN, etc., have been used to learn the mapping from the source speaker's acoustic space to the target speaker's acoustic space. However, referring to the state-of-the art techniques, we see that the GMM based voice conversion techniques are most widely used. The goal of such systems is to incorporate the target speaker characteristics in the transformed speech.

In this thesis, we have proposed ANN based technique for voice conversion and have compared the ANN based voice conversion system with that of the GMM based voice conversion system. We have shown that the ANN based voice conversion performs as good as that of the GMM based system. We have also shown that a smoothing technique such as MLPG is not needed with the use of ANN for voice conversion. To further improve the spectral transformation quality of the ANN based system we have proposed three methods, namely, appending deltas, use of a hybrid model and use of contextual features. However, as these techniques depend on the availability of the parallel training data, use of such techniques may not always be feasible. A further limitation to such technique is that the trained model can be used to transform from the trained source speaker to the trained target speaker. Hence, if a new speaker wants to

transform his voice to the target speaker, the speaker has to provide the recorded utterances and train a system. To avoid the need for training data from both the speakers, i.e, the source speaker and the target speaker, we have proposed a technique that captures speaker specific characteristics such that a model can be trained only on the target speaker data. Such a technique allows us to transform any arbitrary source speaker to the target speaker. Incidentally, the proposed method also finds application in building a cross-lingual voice conversion system, where the source speakers' language and the target speakers' language are different.

5.2 Limitations

1. Our focus in this thesis was to improve the spectral transformation performance and hence we have not laid much emphasis on source features transformation. We have used a Gaussian normalized transformation to scale the source speaker pitch frequency taking their mean and variance into account.

2. Assuming that parallel training data is available, typically 30-50 utterances are used in voice conversion. In this thesis, we have designed an approach that captures speaker-specific characteristics i.e, the target speaker. As 50 utterances will not be sufficient to build such a model, we assume that we have a large amount of target speaker training data on which this model could be built.

3. We propose the use of formants in our approach for capturing speaker specific characteristics. These formants were extracted from a well known tool ESPS. Though the study of extracting formants in a robust manner is not complete, we consider the output of ESPS as standard and use them for our work.

4. Vocal Tract Length Normalization (VTLN) is a speaker normalization technique that tries to compensate for the effect of speaker-dependent vocal tract lengths. There are methods of implementing the same in a robust manner, however, for our experimentation we use a simpler method.

5. Our current approach for CLVC needs a large amount of speech data which would be equivalent to the one needed to build a TTS, however, we have not conducted experiments to find out what could be the optimal size of the data needed to get an acceptable level of performance.

5.3 Future work

- The transformation of spectral features and average pitch frequency is not enough to obtain a good voice transformation. Duration and pitch contours are also a few of the important features that affect the transformation performance.

- The quality of the current cross-lingual voice conversion depends on the accuracy of formant prediction. Most of the current formant extraction techniques are not robust. We are currently using a well known tool ESPS to extract the formants. However as we have not validated the accuracy of it, we intend to do the same and come up with an approach which would be more robust in extracting formants. Theoretically speaking, the number of formants vary from phone to phone, however for our current experiments we use 6 formants for every phone. Hence, one could design an algorithm by carefully considering the nature of the phone and the number of formants.

- In our approach for cross-lingual voice conversion, as we did not use a bilingual speaker, we did not find any means to perform an objective evaluation. Hence, there is a need to come up with an algorithm that can be used to assess the quality of transformation objectively.

REFERENCES

[1] G. Fant, *Acoustic theory of speech production*, Mouton De Gruyter, 1970.

[2] B. Yegnanarayana, *Artificial Neural Networks*, Prentice Hall of India, 2004.

[3] J. Makhoul, "Linear prediction: A tutorial review," in *Proceedings of the IEEE*, 1975, vol. 63, pp. 561–580.

[4] F. Itakura, "Minimum prediction residual principle applied to speech recognition," in *IEEE Trans. Acoust., Speech, Signal Processing*, 1975, vol. 23, pp. 67–72.

[5] W. Holmes, J. Holmes, and M. Judd, "Extension of the bandwidth of the jsru parallel-formant synthesizer for high quality synthesis of male and female speech," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1990, vol. 1, pp. 313–316.

[6] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *IEEE Trans. Acoust., Speech, Signal Processing*, 1990, vol. 28, pp. 357–366.

[7] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1988, vol. 1, pp. 655–658.

[8] E. K. Kim, S. Lee, and Y. H. Oh, "Hidden markov model based voice conversion using dynamic characteristics of speaker," in *European Conference On Speech Communication And Technology*, 1997, pp. 1311–1314.

[9] H. Mori and H. Kasuya, "Speaker conversion in arx-based source-formant type speech synthesis," in *European Conference On Speech Communication And Technology*, 2003, pp. 2421–2424.

[10] H. Duxans, *Voice Conversion applied to Text-to-Speech systems*, PhD dissertation, Universitat Politecnica de Catalunya, Barcelona, 2006.

[11] A Kain and M. W. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelop mapping and residual prediction," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2001, vol. 2, pp. 813–816.

[12] Y. Stylianou, O. Cappe, and E. Moulines, "Statistical methods for voice quality transformation," in *Eurospeech*, 1995, pp. 447–450.

[13] A. R. Toth and A. W. Black, "Using articulatory position data in voice transformation," in *Workshop on Speech Synthesis*, 2007, pp. 182–187.

[14] T. Toda, A. W. Black, and K. Tokuda, "Acoustic-to-articulatory inversion mapping with gaussian mixture model,," in *Proceedings of Int. Conf. Spoken Language Processing*, 2004.

[15] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks,," *Speech Communication*, vol. 16, pp. 207–216, 1995.

[16] T. Watanabe, T. Murakami, M. Namba, T. Hoya, and Y. Ishida, "Transformation of spectral envelope for voice conversion based on radial basis function networks," in *Proceedings of Int. Conf. Spoken Language Processing*, 2002.

[17] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2009.

[18] H. Valaree, E. Moulines, and J. P. Tubach, "Voice transformation using psola technique," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1992.

[19] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, and S. Narayanan, "Text-independent voice conversion based on unit selection," .

[20] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on gaussian mixture model," in *Proceedings of INTERSPEECH*, 2006.

[21] T. Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2007.

[22] A. Kain, *High resolution voice transformation*, PhD dissertation, Oregon Health and Science University, 2001.

[23] H. Ye and S. Young, "Voice conversion for unknown speakers," in *Proceedings of Int. Conf. Spoken Language Processing*.

[24] A. Sundermann, A. Bonafonte, H. Hoge, and H. Ney, "Voice conversion using exclusively unaligned training data," in *ACL/EMNLP*.

[25] A. Mouchtaris, J. V. Spiegel, and P. Mueller, "Non-parallel training for voice conversion by maximum likelihood constrained adaptation," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*.

[26] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm based speech synthesis," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2000.

[27] O. Turk, *Cross-lingual voice conversion*, PhD dissertation, Bogazii University, 2007.

[28] G. Zuo, W. Liu, and X. Ruan, "Genetic neural networks based rbf neural network for voice conversion," in *World congress on intelligent control and automation*.

[29] K. Sreenivasa Rao, "Voice conversion by mapping the speaker-specific features using pitch synchronous approach," *Elsevier Science*, 2009.

[30] D. Erro, *Intra-lingual and cross-lingual voice conversion using harmonic plus stochastic models*, PhD dissertation, Universitat Politcnica de Catalunya, 2008.

[31] H. Matsumoto, S. Hiki, T. Sone, and T. Nimura, "Multidimensional representation of personal quality of vowels and its acoustical correlates," *IEEE Transactions AU*, pp. 428–436, 1973.

[32] K. Itoh and S. Saito, "Effects of acoustical feature parameters of speech on perceptual identification of speaker," *IECE Transactions*, pp. 101–108, 1982.

[33] H. Kuwabara and Y. Sagisaka, "Acoustic characteristics of speaker individuality: Control and conversion," *Speech Communication*, vol. 16, 1995.

[34] K. Shikano, S. Nakamura, and M. Abe, "Speaker adaptation and voice conversion by codebook mapping," in *IEEE International Symposium on Circuits and Systems,*, vol. 1.

[35] L. M Arslan, "Speaker transformation algorithm using segmental codebooks (stasc)," *Speech Communication*, vol. 28.

[36] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1998, vol. 1, pp. 285–288.

[37] A. Pozo, *Voice Source and Duration Modelling for Voice Conversion and Speech*, PhD dissertation, University of Cambridge, 2008.

[38] T. Toda, A. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2005, vol. 1, pp. 9–12.

[39] T. Toda, *High-Quality and Flexible Speech Synthesis with Segment Selection and Voice Conversion*, PhD dissertation, Nara Institute of Science and Technology, 2003.

[40] O. Turk, "New methods for voice conversion," MS dissertation, Boazii University, 2003.

[41] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, and J. Hirschberg, "Text-independent cross-language voice conversion," in *Proceedings of Int. Conf. Spoken Language Processing*.

[42] A. Sundermann, *Text-Independent Voice Conversion*, PhD dissertation, Universitat der Bundeswehr Munchen, 2007.

[43] A. Mouchtaris, J. V. Spiegel, and P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, 2006.

[44] D. Sundermann, H. Ney, and H. Hoge, "Vtln based cross-language voice conversion," .

[45] K. S. Lee, D. H. Youn, and I. W. Cha, "A new voice transformation method based on both linear and nonlinear prediction analysis," in *Proceedings of Int. Conf. Spoken Language Processing*, pp. 1401–1404.

[46] H. Ye and S. Young, "High quality voice morphing," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*.

[47] A. Sundermann, A. Bonafonte, H. Ney, and H. Hoge, "A study on residual prediction techniques for voice conversion," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*.

[48] K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through phoneme based linear mapping functions with straight for mandarin," in *4th International Conference on Fuzzy Systems and Knowledge Discovery*.

[49] O. Turk and L. M. Arslan, "Voice conversion methods for vocal tract and pitch contour modification," in *European Conference on Speech Communication and Technology*, pp. 2845–2848.

[50] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Text-to-speech synthesis with arbitrary speaker's voice from average voice," in *European Conference On Speech Communication And Technology*, 2001.

[51] T. Ceyssens, W. Verhelst, and P. Wambacq, "On the construction of a pitch conversion system," in *Europen Signal Processing Conference*, 2002.

[52] P. D. Auero, J. Adell, and A. Bonafonte, "Improving tts quality using pitch contour information of source speaker in s2st framework," in *International Workshop Advances in Speech Technology*, 2002.

[53] H. Ye and S. Young, "Perceptually weighted linear transformations for voice conversion," in *Proceedings of EUROSPEECH*.

[54] T. Toda, A. W. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis," in *Proceedings of 5th ISCA Speech Synthesis Workshop*, Pittsburg, PA., June 2004, pp. 31–36.

[55] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*.

[56] T. Dutoit, *An introduction to text-to-speech synthesis*, Kluwer Academic Publishers, 1997.

[57] A Kain and M. W. Macon, "Personalizing a speech synthesizer by voice adaptation," in *3rd ECSA/COCOSDA International Speech Synthesis Workshop*, 1998, pp. 225–230.

[58] W. Zhang, L. Q. Shen, and D. Tang, "Voice conversion based on acoustic feature transformation," in *6th national conference on Man-machine speech communications*, 2001.

[59] E. V. Raghavendra, S. Desai, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Global syllable set for speech synthesis in indian languages," in *IEEE 2008 workshop on Spoken Language Technologies, Goa, India*.

[60] H. Hoge, "Project proposal tc-star - make speech-to-speech translation real," in *LREC*, 2002.

[61] O. Turk and L. Arslan, "Subband based voice conversion," in *ICSLP*, 2002.

[62] J. Hosom, A. Kain, T. Mishra, J. V. Santen, M. Fried-Oken, and J. Staehely, "Intelligibility of modifications to dysarthic speech," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2003.

[63] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech," in *Proceedings of INTERSPEECH*, 2006.

[64] Q. Jin, A. Toth, T. Schultz, and A. W. Black, "Voice convergin: Speaker de-identification by voice transformation," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*.

[65] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," in *5th ISCA Speech Synthesis Workshop*.

[66] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*.

[67] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech and Language Processing*, pp. 2222–2235, 2007.

[68] A. W. Black and K. Lenzo, "Building voices in the festival speech synthesis system," in http://festvox.org/bsv/.

[69] S. Furui, "Cepstral analysis for automatic speaker verification," *IEEE Trans. on Audio Speech and Signal Proc*, 1981.

[70] H. Misra, S. Ikbal, and B. Yegnanarayana, "Speaker-specific mapping for text-independent speaker recognition," *Speech Communication*, 2003.

[71] ESPS, "Esps source code from the esps/waves+ package," in *[Online]*.

[72] A. Faria, "Pitch based vocal tract length normalization," in *Tech. Rep. TR-03-001, International Computer Science Institute*.

LIST OF PUBLICATIONS

The work done during my masters has been disseminated to the following journal and conferences.

Journal:

1. Srinivas Desai, B. Yagnanarayana, Alan W Black, Kishore Prahallad, ”Spectral Mapping Using Artificial Neural Networks for Voice Conversion”, *IEEE Transactions on Audio, Speech and Language Processing, 2010.*

Conferences:

1. Srinivas Desai, B. Yegnanarayana, Kishore Prahallad, ”A Framework for Cross-Lingual Voice Conversion using Artificial Neural Networks”, in Proceedings of *International Conference on Natural Language Processing (ICON)*, Hyderabad, India, December 2009.

2. Srinivas Desai, E. Veera Raghavendra, B. Yegnanarayana, Alan W Black, Kishore Prahallad, ”Voice Conversion Using Artificial Neural Networks”, in Proceedings of *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, April 2009.

3. E. Veera Raghavendra, Srinivas Desai, B. Yegnanarayana, Alan W Black, Kishore Prahallad, ”Blizzard 2008: Experiments on Unit Size for Unit Selection Speech Synthesis”, in *Blizzard Challenge 2008 workshop*, Brisbane, Australia, September 2008.

4. E. Veera Raghavendra, Srinivas Desai, B. Yegnanarayana, Alan W Black, Kishore Prahallad, ”Global Syllable Set for Building Speech Synthesis in Indian Languages”, in Proceedings of *IEEE 2008 workshop on Spoken Language Technologies*, Goa, India, December 2008.

CURRICULUM VITAE

1. **NAME:** Srinivas Desai

2. **DATE OF BIRTH:** 06 April 1983

3. **PERMANENT ADDRESS:**

   Srinivas Desai

   S/O Hanumanth Rao Desai

   Plot No: 101,

   Sapthagiri Colony, Sainikpuri,

   Secunderabad - 500094.

   Andhra Pradesh,

   India.

4. **EDUCATIONAL QUALIFICATION:**

   - July 2005: Bachelor of Engineering (ECE), Poojya Doddappa Appa College of Engineering, Gulbarga, Karnataka, India.

THESIS COMMITTEE

1. **GUIDE:** Mr. Kishore Prahallad

2. **MEMBERS:**

    - Prof. Jayanthi Sivaswamy
    - Dr. Garimella Rama Murthy