Voice Conversion:
Alignment and Mapping Perspective

by

Nirmesh J. Shah
201321009

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY
to
DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

December 2019

Declaration

I hereby declare that

i) the thesis comprises of my original work towards the degree of Doctor of Philosophy at Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT) and has not been submitted elsewhere for a degree,

ii) due acknowledgment has been made in the text to all the reference material used.

Nirmesh J. Shah

Certificate

This is to certify that the thesis work entitled, "*Voice Conversion: Alignment and Mapping Perspective*," has been carried out by *Nirmesh J. Shah* for the degree of *Doctor of Philosophy* at *Dhirubhai Ambani Institute of Information and Communication Technology* (DA-IICT) under my supervision.

Prof. (Dr.) Hemant A. Patil
Thesis Supervisor

Acknowledgments

I take this opportunity to extend my sincere gratitude to all those who helped me directly and indirectly for making this Ph.D. thesis possible. First and foremost, I would like to express my sincere gratitude to my Ph.D. supervisor, Prof. (Dr.) Hemant A. Patil, for his dedicated help, encouragement, and continuous support throughout my Ph.D. His passion and dedication for publishing high-quality research work in top conferences and peer-reviewed journals, has made a deep impression on me. During the last seven years, I have learned extensively from him including how to approach a research problem by systematic thinking, how to raise a new possibility, how to face failures and never losing hope. I owe lots of gratitude to him for giving me his time. He always gave utmost priority to my manuscripts and Ph.D. thesis. He supported me as a friend and family member. In particular, he always makes sure that I do not have any personal problem in my life. In particular, he helped me financially throughout my Ph.D, which indeed helped me for doing my research with peace of mind. I am extremely glad to be associated with a person like him in my life.

I sincerely thank my thesis eamination committee members, Prof. B. Yegnanarayana (from IIIT Hyderabad) and Prof. Junichi Yamagishi (National Institute of Informatics, Japan), for the detail thesis report. I am also thankful to Research Progress Seminar (RPS) committee members, namely, Prof. (Dr.) M. V. Joshi and Prof. (Dr.) Sourish Dasgupta for patiently listening to my research progress at the end of each semester and providing me valuable feedback. I also thank my Ph.D. synopsis committee members, Prof. (Dr.) Rajeeb Lochan Das for his feedback. In addition, we thank Prof. (Dr.) Douglas O'Shaughnessy (Professor at INRS-Telecommunications, University of Quebec, Montreal, Canada) and Dr. S. H. Mohammadi (Research Scientist at ObEN inc., USA, and Ph.D. in Voice Conversion from the Center for Spoken Language Understanding (CSLU), Oregon Health & Science University (OHSU)) for their suggestions and kind help in doing English language-related corrections in my Computer, Speech, & Language, Elsevier, manuscript.

I sincerely acknowledge Ministry of Electronics and Information Technology (MeitY), Govt. of India and authorities of DA-IICT, Gandhinagar for their kind help to carry out the research work. I am also thankful to the Resource Center (RC) staff of DA-IICT for their prompt cooperations.

My heartfelt thanks to all my co-authors, Mihir Parmar, Neil Shah, Dr. Hardik B. Sailor, Dr. Maulik Madhavi, Maitreya Patel, Sushant V. Rao, Sreeraj, Mohammadi Zaki, Avni Rajpal, Pramod Bachhav, Bhavik Vachhani, Savan Doshi, Mirali Purohit, Jui Shah, Satyam Kumar, Madhu Kamble, Prasad Tapkir, Meet Soni, for their help, support and technical discussions for doing an excellent teamwork.

I am grateful to my dear friends, Dr. Hardik B. Sailor, Mihir Parmar, Mohammadi Zaki for helping out when I got stuck in any technical difficulties. They always patiently listened to me and provided me a different perspective to my research problem. I also acknowledge Sumukh Bansal for technical discussion with me. My special regards to all my seniors and Ph.D. colleagues, Dr. Maulik Madhavi, Dr. Tanvina B. Patel, Dr. Anshu Chittora, Dr. Hardik B. Sailor, Madhu Kamble, and Ankur Patil for their guidance and moral support. I also thank all my hardworking fellow labmates from Speech Research Lab, DA-IICT.

A special mention of thanks to my friends, Parth, Sumukh, Hardik, Zaki, Anusha, Deshna, Kamal, Gaurav, Sanjay, Anand, Mayur, Pratik, Pramod, Purvi, Avni, Rishabh, Apeksha, Milind Bhaiya who made my Ph.D. journey memorable at DA-IICT Gandhinagar.

I feel a deep sense of gratitude for my father (Mr. Jayesh R. Shah) and mother (Mrs. Jayshree J. Shah), who formed part of my vision. Their infallible love and blessings have always been my strength. Their patience and sacrifice will remain my inspiration throughout my life. I am also very much grateful to my family members Ruchi, Samir, Urja and Dhyana for their love and support. Finally, I would also like to thank the All Mighty GOD, and my teacher Imdad Ali Nandoilya for their continuous blessings.

Nirmesh Jayeshkumar Shah

Contents

Abstract

Understanding how a particular speaker is producing speech, and mimicking one‘s voice is a difficult research problem due to the sophisticated mechanism involved in speech production. Voice Conversion (VC) is a technique that modifies the perceived speaker identity in a given speech utterance from a source speaker to a particular target speaker without changing the linguistic content. Each standalone VC system building consists of two stages, namely, training and testing. First, speaker-dependent features are extracted from both speakers‘ training data. These features are first time aligned and corresponding pairs are obtained. Then a mapping function is learned among these aligned feature-pairs. Once the training step is done, during the testing stage, features are extracted from the source speaker‘s held out data. These features are converted using the mapping function. The converted features are then passed through the vocoder that will produce a converted voice. Hence, there are primarily three components of the stand-alone VC system building, namely, the alignment step, the mapping function, and the speech analysis/synthesis framework.

Major contributions of this thesis are towards identifying the limitations of existing techniques, improving it, and developing new approaches for the mapping, and alignment stages of the VC. In particular, a novel Amplitude Scaling (AS) method is proposed for frequency warping (FW)-based VC, which linearly transfers the amplitude of the frequency-warped spectrum using the knowledge of a Gaussian Mixture Model (GMM)-based converted spectrum without adding any spurious peaks. To overcome the issue of overfitting in Deep Neural Network (DNN)-based VC, the idea of pre-training is popular. However, this pre-training is time-consuming, and requires a separate network to learn the parameters of the network. Hence, whether this additional pre-training step could be avoided by using recent advances in deep learning is investigated in this thesis. The ability of Generative Adversarial Network (GAN) in estimating probability density function (*pdf*) for generating the realistic samples corresponding to the given source speaker‘s utterance resulted in a significant performance improvement in the area of VC. The key limitation of the vanilla GAN-based system is in generating the

samples that may *not* correspond to the given source speaker‘s utterance. To address this issue, Minimum Mean Squared Error (MMSE) regularized GAN (i.e., MMSE-GAN) is proposed in this thesis.

Obtaining corresponding feature pairs in the context of both parallel as well as non-parallel VC is a challenging task. In this thesis, the strengths and limitations of the different existing alignment strategies are identified, and new alignment strategies are proposed for both parallel and non-parallel VC task. Wrongly aligned pairs will affect the learning of the *mapping* function, which in turn will deteriorate the quality of the converted voices. In order to remove such wrongly aligned pairs from the training data, outlier removal-based pre-processing technique is proposed for the parallel VC. In the case of non-parallel VC, theoretical convergence proof is developed for the popular alignment technique, namely, Iterative combination of a Nearest Neighbor search step and a Conversion step Alignment (INCA). In addition, the use of dynamic features along with static features to calculate the Nearest Neighbor (NN) aligned pairs in the existing INCA, and Temporal context (TC) INCA is also proposed. Furthermore, a novel distance metric is learned for the NN-based search strategies, as Euclidean distance may not correlate well with the perceptual distance. Moreover, computationally simple Spectral Transition Measure (STM)-based phone alignment technique that does not require any apriori training data is also proposed for the non-parallel VC.

Both the parallel and the non-parallel alignment techniques will generate one-to-many and many-to-one feature pairs. These one-to-many and many-to-one pairs will affect the learning of the mapping function and result in the *muffling* and *oversmoothing* effect in VC. Hence, unsupervised Vocal Tract Length Normalization (VTLN) posteriorgram, and novel inter mixture weighted GMM Posteriorgram as a speaker-independent representation in the two-stage mapping network is proposed in order to avoid the alignment step from the VC framework. In this thesis, an attempt has also been made to use the acoustic-to-articulatory inversion (AAI) technique for the quality assessment of the voice converted speech. Lastly, the proposed MMSE-GAN architecture is extended in the form of Discover GAN (i.e., MMSE DiscoGAN) for the cross-domain VC applications (w.r.t. attributes of the speech production mechanism), namely, Non-Audible Murmur (NAM)-to-WHiSPer (NAM2WHSP) speech conversion, and WHiSPer-to-SPeeCH (WHSP2SPCH) conversion. Finally, thesis summarizes overall work presented, limitations of various approaches along with future research directions.

List of Acronyms

AAI	Acoustic-to-Articulatory Inversion
AE	AutoEncoder
ANF	Auditory Nerve Fiber
ANN	Artificial Neural Network
AS	Amplitude Scaling
ASR	Automatic Speech Recognition
ASV	Automatic Speaker Verification
BLFW	Bi-Linear Frequency Warping
BUT	Brno University of Technology
CC	Correlation Coefficient
CHAINS	CHAracterizing INdividual Speakers
COVAREAP	COoperative Voice Analysis REPository for speech technologies
DAE	Denoising AutoEncoder
dim	Dimensional
DiscoGAN	Discover Generative Adversarial Network
DNN	Deep Neural Network
DTW	Dynamic Time Warping
EE	Estimation Error
ELU	Exponential Linear Unit
EM	Expectation Maximization
EMA	ElectroMagnetic Articulography
FMcovD	Fast Minimum Covariance Determinant
FW	Frequency Warping
GAN	Generative Adversarial Network
GMM	Gaussian Mixture Model
GP	Gaussian Posteriorgram

GSC	Generalized Smoothness Criterion
GV	Global Variance
HMM	Hidden Markov Model
IMW	Inter Mixture Weighted
INCA	Iterative combination of a Nearest Neighbor search step and a Conversion step Alignment
JDGMM	Joint Density Gaussian Mixture Model
LMNN	Large Margin Nearest Neighbor
LReLU	Leaky Rectifier Linear Unit
MCC	Mel Cepstral Coefficient
MCD	Mel Cepstral Distortion
McovD	Minimum Covariance Determinant
MFCC	Mel Frequency Cepstral Coefficient
MI	Mutual Information
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimation
MMSE	Minimum Mean Squared Error
MOCHA	MultiCHannel Articulatory
MOS	Mean Opinion Score
MT	Machine Translation
MV	Mean Variance
NAM	Non-Audible Murmur
NAM2SPCH	Non-Audible Murmur-to-Speech Conversion
NAM2WHSP	Non-Audible Murmur-to-Whisper Speech Conversion
NN	Nearest Neighbor
OBLFW	Only Bi-Linear Frequency Warped
OCWSS	Open Condenser Wrapped with Soft Silicon
PA	Phonetic Accuracy
PAS	Proposed Amplitude Scaling
PCA	Principal Component Analysis
PCC	Pearson Correlation Coefficient
PESQ	Perceptual Evaluation of Speech Quality
PG	PosteriorGram
PLPCC	Perceptual Linear Prediction Cepstral Coefficients

PPG	Phonetic PosteriorGram
PSD	Positive SemiDefinite
QbE-STD	Query-by-Example Spoken Term Detection
RBM	Restricted Boltzmann Machine
ReLU	Rectifier Linear Unit
RMSE	Root Mean Squared Error
ROBPCA	Robust Principal Component
SD	Standard Deviation
SGD	Stochastic Gradient Descent
SS	Speaker Similarity
SSD	Spoofed Speech Detection
STC	Split-Temporal Context
STM	Spectral Transition Measure
STS	Speech-To-Speech
SVD	Singular Value Decomposition
TC	Temporal Context
t-SNE	*t*-Stochastic Neighbor Embedding
TTS	Text-To-Speech
VAD	Voice Activity Detection
VC	Voice Conversion
VCC	Voice Conversion Challenge
VQ	Vector Quantization
VT	Voice Transformation
VTLN	Vocal Tract Length Normalization
WHSP2SPCH	Whisper-to-Speech Conversion

List of Symbols

$p(\cdot)$	Probability distribution
$P(\cdot)$	Probability of an event
$\mathcal{F}(\cdot)$	Mapping function
$\mathbf{X}$	Source feature vector sequence
$\mathbf{Y}$	Target feature vector sequence
$\mathbf{Z}$	Joint feature vector
λ	Gaussian Mixture Model (GMM) parameters
ω_m	Weight of the m^{th} mixture component
μ_m	Mean of the m^{th} mixture component
Σ_m	Covariance of the m^{th} mixture component
N_c	Number of mixture components
α	Warping factor
W_α	Warping matrix
c	Speed of sound (i.e., 350 m/s)
l	Length of the vocal tract
$p_m^\lambda(x)$	Probability of x belongs to m^{th} mixture component.
$\prod$	Notation for multiplication
$\sum$	Notation for summation
$\mathbf{h}$	Hidden units
$\mathbf{W}$	Weights of a model (DNN, RBM, ConvRBM)
u	Leakage parameter
$\odot$	Elementwise multiplication
$\approx$	Approximately
$\Delta\Delta$	Delta-delta or double-delta or acceleration features
$\langle\cdot\rangle$	Sample average
$\mathbb{E}[\cdot]$	Expectation operator
ϵ	Learning rate parameter

η	Momentum parameter
$\lVert\cdot\rVert_p$	L^p-norm, where $p = 2$ for L^2-norm
$Bernoulli(\cdot)$	Bernoulli distribution
F_0	Fundamental frequency
w	Warping path
$\mathbf{1}_n$	Column vector with all n components having the value 1
χ^2	Chi-squared distribution
$\mathcal{S}$	Set of similar pairs
$\mathcal{D}$	Set of dissimilar pairs
C_k	k^{th} cluster
$P(C_k \vert O_t)$	Posterior probability of the current frame o_t for the k^{th} cluster
$\mathbf{X}^\alpha$	Warped feature sequence
$P(\mathbf{X} \vert \lambda)$	Likelihood value of $\mathbf{X}$
li	Articulatory parameters for lower incisor
ul	Articulatory parameters for upper lip
ll	Articulatory parameters for lower lip
tt	Articulatory parameters for tongue tip
tb	Articulatory parameters for tongue body
td	Articulatory parameters for tongue dorsum
v	Articulatory parameters for velum
ui	Articulatory parameters for upper incisor
bn	Articulatory parameters for bridge of the nose
I	Mutual Information
$S(z)$	$\mathcal{Z}$-transform of discrete-time speech signal
$W(z)$	$\mathcal{Z}$-transform of discrete-time whispered speech signal
$N(z)$	$\mathcal{Z}$-transform of discrete-time NAM signal
$V(z)$	Vocal tract transfer function from the glottis to the lips
$R(z)$	System function for the lip radiation
$H(z)$	$\mathcal{Z}$-transform of the impulse response of the NAM microphone
$G(z)$	$\mathcal{Z}$-transform of the glottal flow input
A_v	Gain corresponding to the loudness of normal speech
A_w	Gain corresponding to the loudness of whispered speech
A_n	Gain corresponding to the loudness of NAM signal
$\succeq$	To indicate positive semidefinite matrix

List of Tables

List of Figures

Chapter 1
Introduction

1.1 Introduction

Understanding how humans produce a speech is considered one of the most challenging tasks due to the sophisticated mechanism involved in speech production [35]. On top of this, understanding how a particular speaker is producing speech and mimicking his/her voice will be all the more difficult research problem. Fewer attempts have been made in the past to actually understand how the mimicry is performed from speech production viewpoint [36, 37]. Voice Conversion (VC) is a technique that modifies the perceived speaker identity in a given speech utterance from a source speaker to a particular target speaker without changing the linguistic content of the utterance [16]. Basically, it can be considered as a speaker conversion technique. In particular, the goal of the VC technique is to mimic the given target speaker similar to professional human mimicry.

Trying to produce speech like someone else, either by mimicking, or by synthesizing, or by a vice conversion is a challenging task, as it requires understanding of speaker-specific characteristics in speech. The speaker-specific charesteristics vary with language, contet and environment. Moreover, the acceptibility of the resulting speech depends on the perception of the listener, which in turn depends on his/er background. VC techniques come under the broad area of Voice Transformation (VT). VT can be considered as any non-linguistic modifications one may apply to the speech signal [15]. For example, time-scaling, pitch-scaling, voice-individuality control, speaker identity conversion, etc. An excellent survey article for the VT can be found in [15]. Unlike speaker recognition/verification, the objective in the VC is not just to identify or detect the speaker-dependent signatures in the speech signal, rather modify these speaker-dependent signatures from one speaker to the another, and to generate high quality converted speech at the end. Hence, VC can be considered as one of the most difficult research issues among all the possible VT techniques.

Depending on the nature of training data, VC can be broadly categorized into parallel and non-parallel cases [14]. In the parallel case, both the speakers have spoken the same utterances from the same language. On the other hand, in the non-parallel case, both speakers have spoken the different utterances from the same or different languages. Each standalone VC system building consists of two stages, namely, training and testing as shown in Figure 1.1. First, speaker-dependent features are extracted from both the speakers' training data. These features are first time aligned and corresponding pairs are obtained. Then, a mapping function is learned from these time-aligned corresponding feature-pairs. Once the training step is completed, during the testing stage, features are extracted from the source speaker's held out data. These features are then converted using the conversion function of the learned mapping function (during training). The converted features are then passed through the vocoder that will produce a converted voice. Hence, there are primarily three components of the VC system building, namely, the alignment step, the mapping function, and the speech analysis-synthesis framework. Next, discussions related to the key potential research challenges associated with the VC problem are presented followed by contributions from this thesis work.

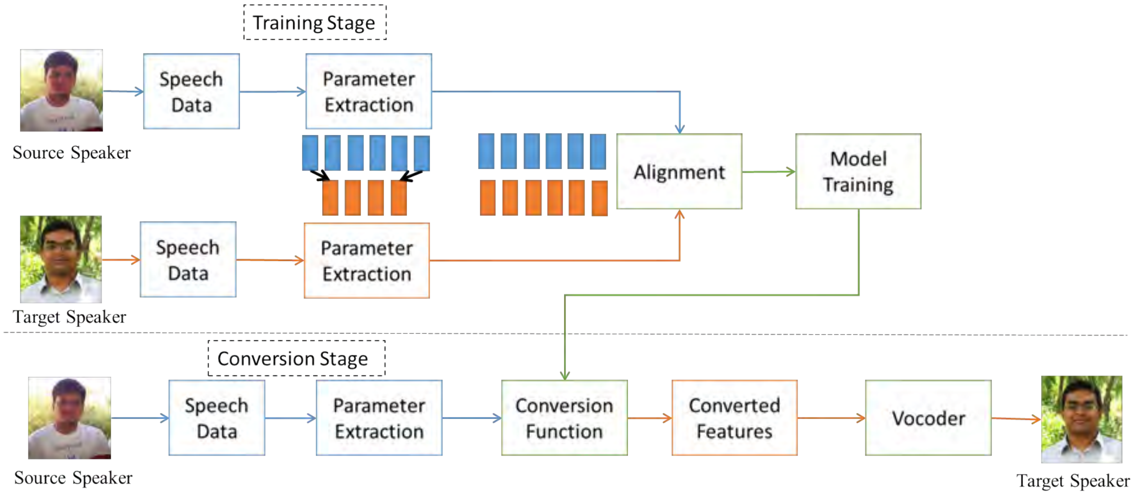


Figure 1.1: Overall system architecture of standalone VC. After [14].

1.2 Motivation

VC is a highly active research area. Numerous VC techniques have been developed over the last four decades. In particular, research focus in the area of VC has increased significantly for last one decade with the advent of deep learning, and the high-quality vocoders. The VC area is still open research problem as 100%

effective system is yet to be developed. Apart from the challenges involved in the VC, one of the key reasons to have significant attention in the speech research community is the interesting applications of VC.

- **Personalized Text-to-Speech (TTS)**: Developing high-quality TTS for a particular speaker requires a large amount of speech corpus. VC techniques can be used to generate different voices with the less amount of available training data from a target speaker to develop personalized TTS system in different voices.

- **Speech-To-Speech (STS) Translation**: STS can convert speech spoken in one language to another language. For example, let‘s consider English-to-Gujarati STS translation, where Gujarati is one of the official languages of India. This is primarily achieved by combining Automatic Speech Recognition (ASR) developed on English database, Machine Translation (MT) for English-to-Gujarati text conversion, and TTS developed on Gujarati database. The key issue with this kind of approach is that at the end, converted voice in the Gujarati language will be perceived as if it has been spoken by the speaker whose voice is used to develop Gujarati TTS. To make conversation across language more natural, converted voices should have a voice similar to the speaker who has spoken in the English language. VC techniques can be used to make STS more natural.

- **Voice Pathology**: The patients that are suffering from the vocal fold-related disorders, such as vocal fold paralysis, vocal nodule, etc. may not be able to produce normal audible speech, due to the partial or complete absence of the vibrations of the vocal folds [38]. This results in glottis being open during the phonation. Hence, the noisy excitation source will be generated, which results in a breathy or whispered voice. Losing a natural way of producing speech affects tremendously one's life, since speech is the most powerful and natural mode of communication among humans. VC techniques can be used to enhance the noisy low-power whispered speech into normal speech.

- **Movie Dubbing**: Movie industry is one of the biggest revenue generating industries. Now, most of the movies are converted into more than one language. Professional voice over artists are hired for the dubbing of movie from one language to another language. Their voices do not sound similar to the original movie actor. VC techniques can convert the dubbing artists' voice into the original movie actor's voice. Hence, dubbing of movie in different languages with the application of VC will generate a perception that as if the original actor is speaking movie in different languages.

- **Gaming Industries**: VC techniques can be used to generate novel multiple voices, which can be used by the gaming or other entertainment industries.

- **Robust Voice Biometric Authentication Systems**: Voice-based biometric authentication system has now become available in the various places with the remarkable success obtained by the speech research community. However, these voice-based biometric authentication systems are vulnerable to various spoofing attacks. Among all possible spoofing attacks, VC pose a great threat to the Automatic Speaker Verification (ASV) systems. Hence, understanding in detail the VC techniques will definitely help in designing spoofing countermeasure for the robust voice biometric authentication system.

- **Data Augmentation strategies**: With the advent of deep learning, notable success have been reported in the various speech research problems. However, deep learning-based approaches work well in the presence of a huge amount of speech database. Obtaining and preparing such huge amount of data will be costly in most of the applications. Hence, VC can be used to generate huge speech data with different speaker characteristics for numerous speech applications.

Apart from the above-mentioned applications, VC mapping techniques can also be applied to other important speech research problems, such as narrowband-to-wideband speech conversion, whisper-to-normal speech conversion, silent speech-to-normal speech conversion, acoustic-to-articulatory conversion, noisy-to-clean speech conversion (speech enhancement), speech-to-singing conversion, etc.

1.3 Key Research Challenges in Voice Conversion (VC)

The key research challenges in developing the VC systems are as follows:

- Exactly mimicking a particular target speaker with a high quality converted voice is still not up to the mark. One of the possible reasons could be not fully exploiting the knowledge of speech production (such as nonlinear source-filter interaction), and speech perception along with speech prosody, and speaking style.

- Better alignment strategies are required in the case of parallel as well as non-parallel VC. In order to get the accurately aligned source and the target speakers' spectral features, there is a need of the phoneme-level boundaries from both the speakers [39]. However, obtaining the accurate alignment at the phoneme-level is a very challenging problem due to the *co-articulation* of the speech sound units, which leads to splitting or disappearing of a current sound due to interference or merging with the adjacent sound units (due to local *vs*. global co-articulation) [17,40].

- Alignment techniques will generate one-to-many and many-to-one feature pairs (as studied in [6, 39]). In addition, if the word spoken by a source speaker is repeated by the target speaker with different variations, it will generate such pairs. Furthermore, if the same word is repeated for several times, it will result in a different speech pattern. This also generates such kind of pairs. Directly learning the relationship in the presence of such pairs is very challenging.

- Most of the real-world applications of VC suffer from the data scarcity issue due to the availability of less amount of training data from the target speaker. Training the mapping function in the case of less amount of training data may lead to overfitting. Hence, there is a need for improved training algorithms that takes care of the issues related to the overfitting.

- In most of the statistical-based VC techniques, converted features are the result of the statistical averaging that leads to over-smoothing and hence, deteriorates the quality of the converted voice. Better residual compensation techniques are required for the post-processing of the converted voices.

- In the case of Maximum Likelihood (ML)-based optimization, the network numerically optimizes the parameters. However, the reduction in the numerical estimates does not always correlate with the generated sample quality [41]. In addition, the ML-based optimization criteria put prior assumptions on the data distribution (such as, the Minimum Mean Square Error (MMSE) objective function assumes the output variables to be Gaussian), which may not be valid for the given data. Hence, such assumptions prevent the network to learn *perceptually* optimal network parameters.

- Most of the VC techniques convert the voice at the frame-level. These frame-based conversion techniques do not effectively convert the speech at suprasegmental-level. Capturing, and converting such a suprasegmental feature is a very challenging task. Modeling prosody for the statistical-based VC approaches is still an open research problem.

- Evaluation of converted voices primarily relies on subjective tests, which are time-consuming and costly. On the other hand, Mel Cepstral Distortion (MCD) is considered as one of the state-of-the-art objective measures in the VC literature. However, it has been observed that most of the time, MCD does *not* correlate well with the perceptual results, i.e., subjective scores [12, 14, 42–46] in the VC literature. Subjective tests rely solely on sophisticated human perception mechanism for hearing. It is a very formidable task to quantify such complex psychological *cognitive* factors via a simplified mathematical analysis. Hence, better objective evaluation strategies are required for the quality assessment of the converted voices.

1.4 Contributions from the Thesis

Major contributions of this thesis are towards identifying the limitations of existing techniques, improving it, and developing new approaches for the mapping and alignment stages of the VC. In particular, following are the key contributions in this thesis from mapping, and alignment perspectives of VC problem:

- **Mapping Perspectives**:
  - **Novel Amplitude Scaling (AS) Method:** In Frequency Warping (FW)-based VC, a novel AS technique is proposed which linearly transfers the amplitude of the frequency-warped spectrum using the knowledge of a Gaussian Mixture Model (GMM)-based converted spectrum without adding any spurious peaks [22]. The novelty of the proposed approach lies in avoiding a perceptual impression of wrong formant location (due to perfect match assumption between the warped spectrum, and the actual target spectrum in state-of-the-art AS method [42]) leading to deterioration in the quality of converted voices.
  - **Improved training of Deep Neural Network (DNN)-based VC:** DNN-based VC techniques suffer from the issue of overfitting due to less amount of available training data from a target speaker [47]. To alleviate this, pre-training is used for better initialization of the DNN parameters, which leads to faster convergence [47]. However, this pre-training is time-consuming, and requires a separate network to learn

the parameters of the network. To that effect, we propose to exploit recently advanced methods to train DNN without pre-training [8].

  - **Generative Adversarial Networks (GANs):** Minimum Mean Squared Error (MMSE) regularized GAN is proposed due to its ability in estimating density for generating realistic samples corresponding to the given source speaker's utterance [23,41,48].

- **Alignment Perspectives:**

  - **Outliers removal for parallel VC:** Limitations of the alignment strategy in the parallel VC are identified [5,39,49]. To overcome these limitations, Robust Principal Component Analysis (ROBPCA)-based outliers removal technique is proposed as a pre-processing step in VC [5,49].
  - **Alignment strategies for non-parallel VC:** Theoretical convergence proof is developed for the popular alignment technique, namely, Iterative combination of a Nearest Neighbor search step and a Conversion step Alignment (INCA) in the case of non-parallel VC [20]. In addition, we also proposed to use dynamic features along with static features to calculate the Nearest Neighbor (NN) aligned pairs in the existing INCA, and Temporal context (TC) INCA algorithms [7].
  - **Metric learning for a non-parallel VC:** Euclidean distance may not correlate well with the perceptual distance. Hence, we propose to learn distance metric using the Large Margin Nearest Neighbor (LMNN) technique that gives a minimum distance for the same phoneme uttered by the different speakers, and more distance for the different set of phonemes for the alignment task in the non-parallel VC [8].
  - **Phone-aware alignment strategy:** The performance of the Nearest Neighbor (NN)-based alignment techniques improve with the information about phone boundaries. However, estimating the exact phone boundary is a challenging task. We propose to exploit a computationally simple Spectral Transition Measure (STM)-based phone alignment technique that does not require any apriori training data [9,28].
  - **Overcoming the need of alignment in VC:** We propose the unsupervised Vocal Tract Length Normalization (VTLN) posteriorgram, and novel Inter Mixture Weighted (IMW) GMM Posteriorgram as a speaker-independent representation in the two-stage mapping network in order to avoid the alignment step [10,11,31].

- **Quality assessment of converted voices:** The ability of humans to speak effortlessly requires the coordinated movements of various articulators, muscles, etc. This effortless movement contributes towards a naturalness, intelligibility, and speaker's identity (which is partially present in voice converted speech). We propose a novel application of the acoustic-to-articulatory inversion (AAI) towards a quality assessment of the voice converted speech [12].

- **Cross-domain conversion:** We extended the proposed GAN-based mapping techniques for other speech conversion techniques. In particular, we extend the proposed MMSE-GAN architecture in the form of Discover GAN (i.e., MMSE DiscoGAN) for the cross-domain (w.r.t. attributes of the speech production mechanism) conversions in the case of Non-Audible Murmur (NAM)-to-WHiSPer (NAM2WHSP) speech conversion, and WHiSPer-to-SPeeCH (WHSP2SPCH) conversion [24, 50].

1.5 Organization of Thesis

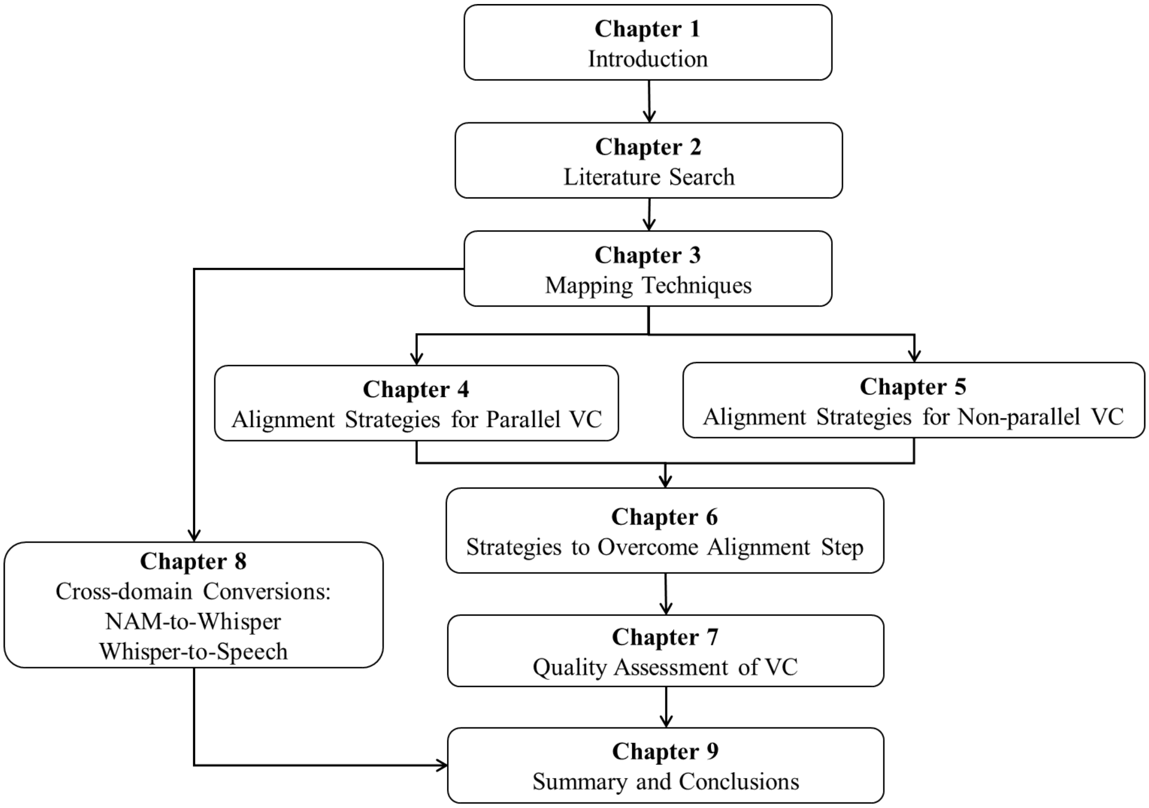


Figure 1.2: Flowchart depicting the organization of the thesis.

The organization of the thesis is shown in Figure 1.2. Chapter 2 briefly summarizes the literature survey of the VC. Chapter 3 presents various state-of-the-art mapping techniques along with the proposed mapping techniques. Chapter 4 and Chapter 5 presents the proposed alignment strategies in the case of parallel and non-parallel VC tasks, respectively. The key limitations of the alignment strategies in both cases have been discussed in Chapter 4 and Chapter 5. Furthermore, two-stage conversion techniques have been proposed in order to overcome the alignment step in both the VC tasks, which is presented in Chapter 6. Chapter 7 discusses the novel application of AAI towards the quality assessment of converted voices. Mapping techniques proposed in Chapter 3 have been applied to the cross-domain conversion applications of VC, which is presented in Chapter 8. Finally, Chapter 9 summarizes the overall work presented in this thesis along with the limitation of the proposed approaches, and future research directions.

1.6 Chapter Summary

In this chapter, the introduction, and motivation for the problem of VC are presented followed by a discussion on its various potential applications along with its basic system architectures. Key research issues or challenges in the VC were discussed along with a brief summary of contributions in the thesis. Finally, the chapter concludes with the organization of the thesis. In the next chapter, the literature search on the VC problem is presented in order to identify gap areas or potential research issues that need immediate attention in the VC problem. This also brings out clearly the key motivations for contributions made in this thesis work.

CHAPTER 2

Literature Survey

2.1 Introduction

In the last chapter, the introduction to the problem of VC was presented. In this chapter, we present history and selected chronological progress in this technologically challenging VC problem. In particular, we present the strengths and weaknesses of various approaches for VC w.r.t. mapping and alignment. This analysis of literature helps in bringing out various research issues or gap area that needs immediate attention in the VC task and thus, the motivation for contributions made in this thesis work. In particular, different stages of VC system building have been discussed in details along with their brief literature search.

2.2 Overview of Voice Conversion

This thesis primarily focuses on the task of Voice Conversion (VC). VC is a technique, which modifies the perceived speaker identity in a given speech signal from a source speaker to a particular target speaker without changing the linguistic content [14]. It is different from Voice Transformation (VT) and voice morphing. In particular, VT refers to any non-linguistic alteration given to the speech signal. In fact, VC is a special type of VT. On the other hand, voice morphing is a technique, which mixes two speakers' voices, and creates a third unknown speaker's voice in order to hide the original speaker's identity for security purposes. The broad categories of VT are shown in Figure 2.1.

Depending on the nature of training data, the problem of VC can be broadly be categorized into two parts, namely, Parallel and Non-Parallel VC (as shown in Figure 2.2) [14]. In parallel VC, both source, and the target speakers have spoken the same utterances. Whereas in non-parallel VC, both source, and target speakers may have spoken different utterances. In addition, utterances spoken by both the speakers may come from the same language or the different languages based on that it can be classified as, intra-lingual or cross-lingual VC, respectively. Furthermore, VC techniques can be classified as text-dependent and text-independent categories. Text-dependent VC needs phonetic transcriptions along with the audio data [51]. On the other hand, text-independent VC systems do not require any phonetic information or transcription [44].

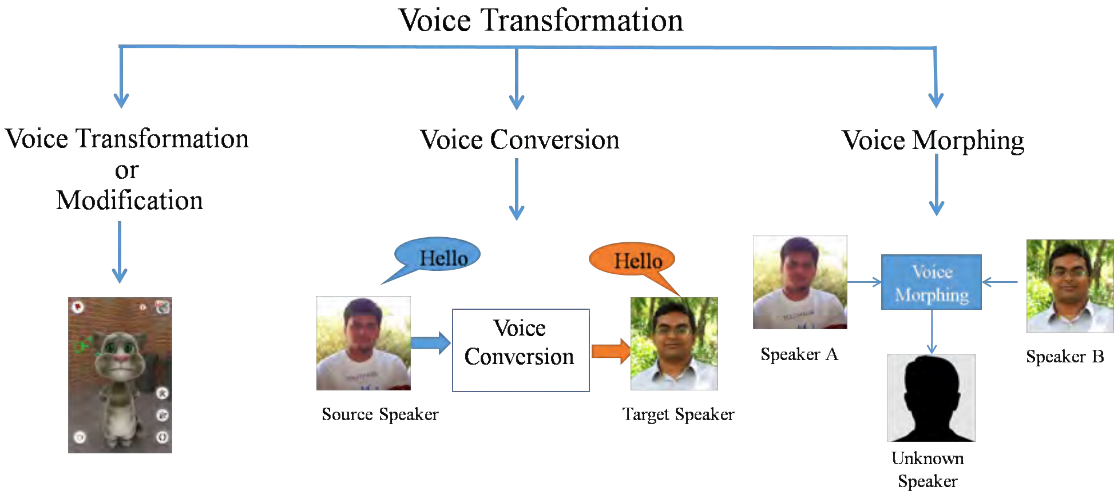


Figure 2.1: Broad categories of Voice Transformation (VT). After [15].

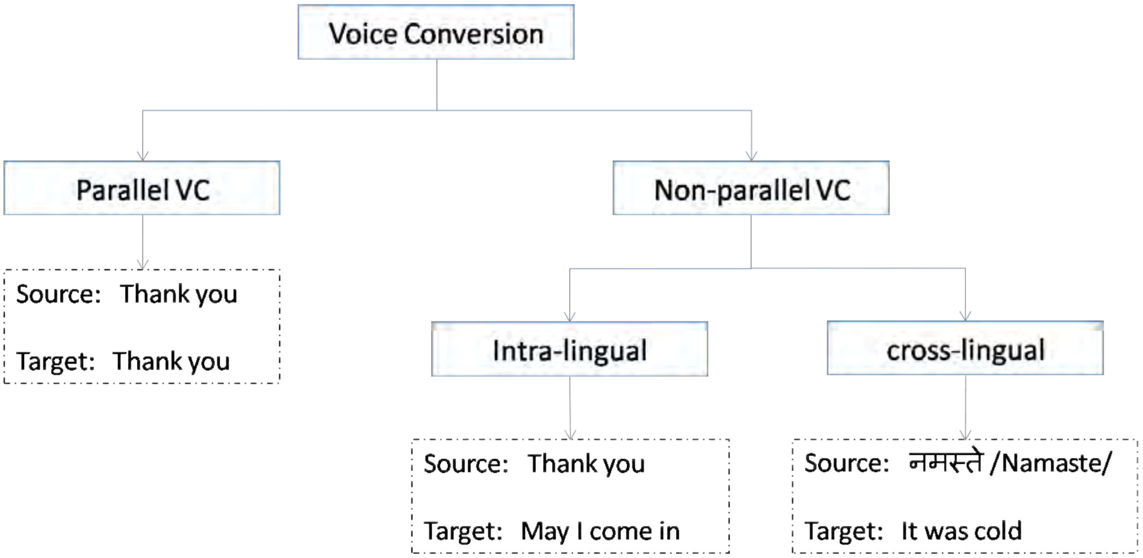


Figure 2.2: Classification of VC-based on nature of training data. After [14]. Here, /Namaste/ means 'Hello' in English.

The conventional VC approaches extract features from the source speaker's speech signal. Then, convert the features from source speaker's representation to the target speaker's representations and at the end, using speech synthesis techniques to generate speech signal from these converted features as shown in Figure

2.3. In the context of VC, we would like to extract, and modify only those features that represent speaker identity, since the speech signal contains different levels of information (such as linguistic, sub-segmental, segmental, supra-segmental, para-linguistic, etc.).

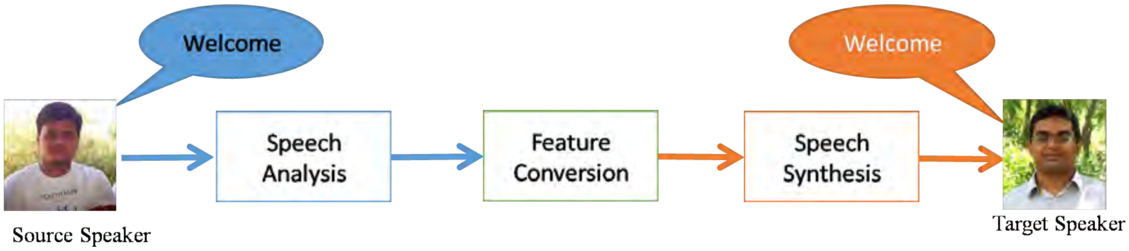


Figure 2.3: Schematic representations of Voice Conversion. After [15].

The conventional VC system building process, regardless of its application, consists of two phases, namely, training and conversion. The block diagram of the stand-alone VC system is shown in Figure 2.4. It consists of three important stages of the VC system building process, namely, speech analysis-synthesis step, alignment step, and mapping step. Each of these steps is discussed in detail along with its chronological progress in the next section.

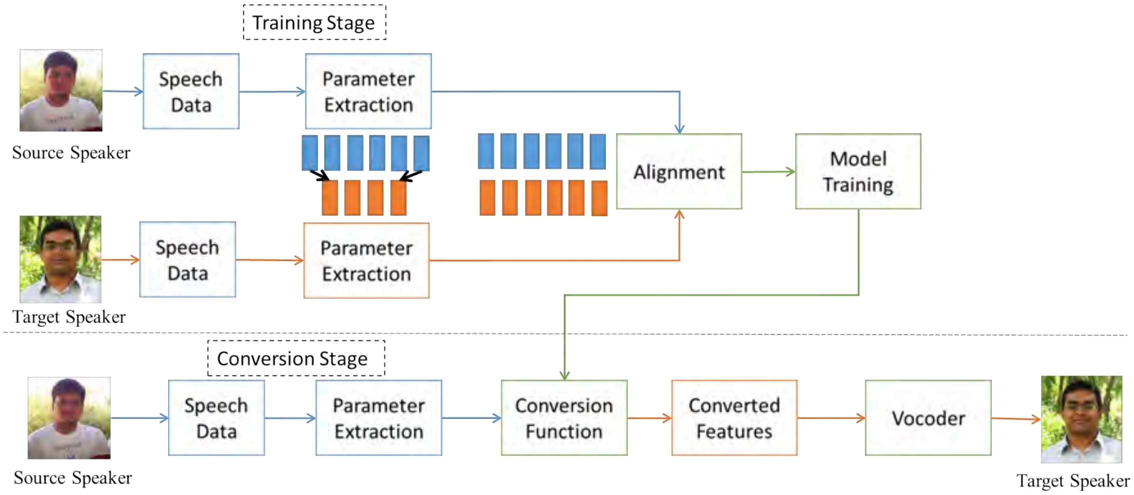


Figure 2.4: Overall system architecture for stand-alone VC. After [16].

2.3 Speech Analysis and Synthesis

Since the task is to convert perceived speaker identity in a given speech signal, one is only interested to extract the features that are responsible for speaker identity, and convert those features from the source speaker's representations to the target speaker’s representations. The factors that are responsible for speaker individuality can be primarily categorized into two parts, namely, sociological *vs.* physiological factors [52]. The speaking style of an individual is more dependent on the sociological factors, such as the community to which the speaker belongs, his/her socio-economic status, dialect, etc. Acoustically, speaking styles can be realized via prosodic features, such as fundamental frequency (F_0) contour, duration, rhythm, power levels, etc. On the other hand, physiological factors are related to the speech organs of the individual speakers. These factors affect the shape and length of the vocal tract system, and hence, the shape of the spectral envelope, spectral tilt, formant frequencies, etc. These sociological and physiological factors can be imagined as software and hardware, respectively. When someone mimics another person, he tries to copy the ’software’ of the target speaker. However, modifying the software part is more challenging in the context of current speech technologies. Hence, most of the VC systems are more concerned with the hardware aspects than the software [15]. Among all the acoustics characteristics, it has been observed in the VC literature that average spectrum, formants, and pitch (F_0) are the most relevant factors for the speaker‘s individuality. Hence, most of the VC techniques try to modify the short-time spectrum and pitch (F_0) [14,52,53]. In particular, the speech analysis-synthesis technique is used to separate the excitation source information from the vocal tract system information, and allows the reconstruction of the speech signal after independently converting both the parameters.

There are primarily three types of speech analysis-synthesis models, namely, source-filter [54, 55], the signal-based [56, 57], and WaveNets [58]. The earlier two models are motivated by the speech production model. During speech production, air rushes from the lungs, and passes through the trachea, and enters the larynx. Then, the vibration of the vocal folds takes place and generates periodic, and noisy airflow for voiced and unvoiced sounds, respectively. This periodic or noisy source excitation then passes through the vocal tract and excites the vocal tract. The vocal tract is the tube-like passage which runs from the glottis at one end and with two openings, oral and nasal cavity, at the other end. The schematic representation of the human speech production system is shown in Figure 2.5. The source-filter model is primarily motivated by the speech-production model. In the source-filter model, speech is modeled as a combination of excitation source (representing the vocal folds), and a spectral envelop filter (representing vocal tract system). This model assumes that the excitation source and vocal tract systems are independent of each other. In particular, this model filters (i.e., convolve w.r.t. the LTI assumption) the excitation source with the vocal tract system to reconstruct the speech signal. Linear Prediction (LP) of Speech and the Mel Log-Spectrum Approximation (MLSA) are the two well-known models used to represent the all-pole and log-spectrum filters for the vocal tract system representations, respectively [54, 59].

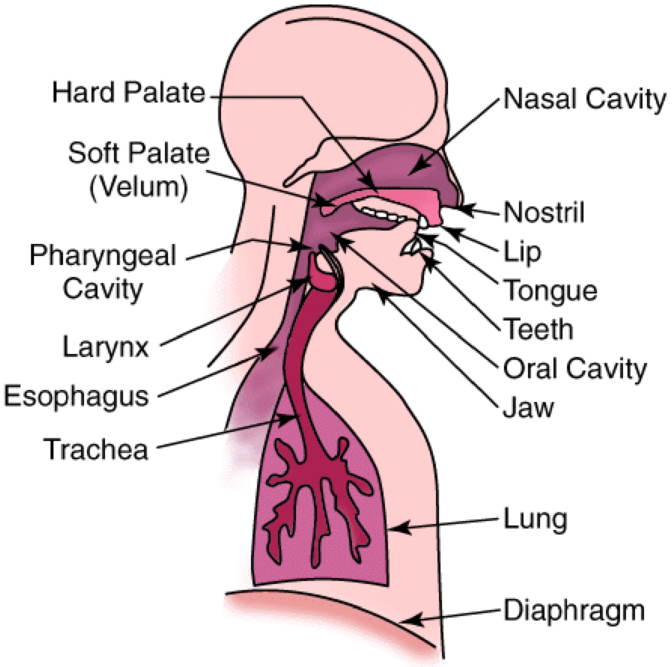


Figure 2.5: Schematic representations of human speech production mechanism. Adapted from [17].

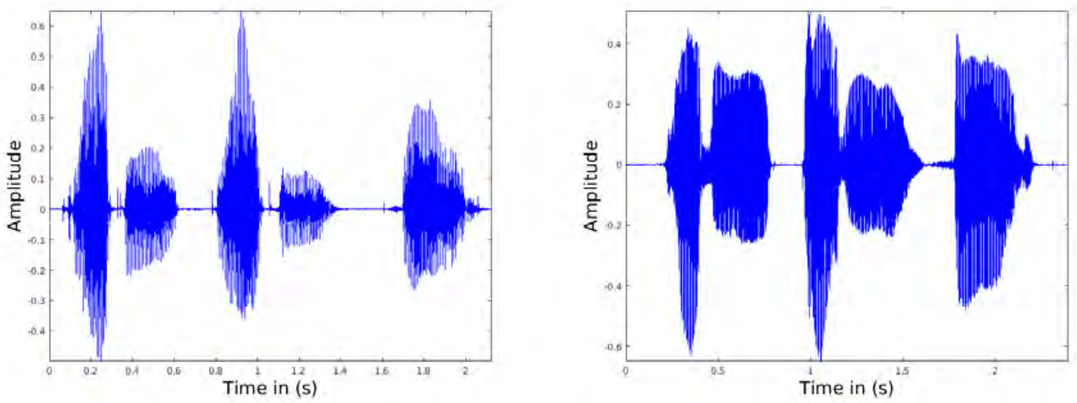
The key limitation of the approach is that the pitch period (i.e., $T_0 = \frac{1}{F_0}$) present in the speech will come as harmonics when estimating the spectral envelope. This is in contradiction with the assumption of *independence* between the excitation source and vocal tract filter. In order to alleviate the interference between the pitch (i.e., F_0) harmonics and the vocal tract spectrum, pitch adaptive time-frequency spectral smoothing technique is proposed in the STRAIGHT vocoder framework [55]. Later on, high-quality TANDEM-STRAIGHT was proposed, which allows unified estimation of the spectrum, fundamental frequency (F_0), and aperiodicity [60]. Recently, CheapTrick and WORLD vocoders reported improvement over the TANDEM-STRAIGHT vocoder [61, 62]. The most common way of modeling excitation source is via the pulse/noise model, where for voiced segments periodic pulses are used, and for the unvoiced segments, noises are used [53]. In addition, more complex excitation source models are also there, for example, glottal excitation model [63], residual signal models [64], mixed excitation [65], and band aperiodicity [66], etc.

Apart from this signal processing, signal-based analysis-synthesis techniques are there, which do not have any independence assumption between the excitation source, and the vocal tract filter. Hence, they yield better performance. However, they are less flexible for the modification. Pitch Synchronous Overlap-Add (PSOLA) [56], Harmonic plus Noise Model (HNM), and its variants are the examples of signal-based approaches [57, 67, 68]. HNM produces high-quality speech and assumes that the speech signal can be decomposed into harmonics, i.e., sinusoid frequencies related to pitch (F_0). In this thesis, AHOCODER is used that produces high-quality speech using the HNM synthesis approach [67].

Recently, WaveNet generative model-based architectures have been proposed [58, 69–71], which is able to produce human-like natural speech signal. WaveNet is a deep autoregressive network, which generates high-fidelity speech waveforms in a sample-by-sample manner [58]. In WaveNet, the Convolutional Neural Network (CNN) takes a raw signal as an input and synthesizes speech an output one sample at a time. It does so by sampling from a softmax distribution of a signal value that is encoded using μ-law transformation and quantized to the 256 possible values. At the time of its release, Deep Mind claimed that WaveNet required too much computational processing power to be used in real-world applications [58]. However, Google announced approximately a 1,000-fold performance improvement along with better voice quality in 2017. WaveNet was then used to generate voices in English and Japanese Google Assistant across all Google platforms [58, 71]. Recently, in the context of VC, it has also become very popular [71].

2.4 Alignment Step for VC

VC is primarily considered as a *supervised* learning problem. However, obtaining corresponding feature pairs among which the mapping has to be learned is a really very challenging task in both parallel and non-parallel VC. Though both the speakers have spoken the same utterances in the case of parallel VC, the spectral features from both source and the target speakers must be time-aligned during training (as shown in Figure 2.6) due to speaking rate variation across the speakers (i.e., interspeaker variations), and speech rate variations within the speaker (i.e., intraspeaker variations). Different time lengths of the same speech utterance spoken by the source and the target speakers will result in a different number of features. Hence, time alignment approaches should be used to compensate for the temporal differences, i.e., to obtain the same number of features from both the speakers. On the other hand in the case of non-parallel VC, utterances spoken by both the speakers will be different. Hence, obtaining corresponding feature pairs in the case of non-parallel VC is the most difficult task. Hence, it must be aligned before applying the standalone VC. Wrongly aligned pairs will affect the learning of the mapping function, which in turn will deteriorate the quality of the converted voices [?, 8, 39, 49, 72]. Hence, alignment is one of the key steps in the VC task.



(a) (b)

Figure 2.6: An example of speech utterance ("Robbery, bribery, fraud") spoken by two different speakers. Taken from the CMU-ARCTIC database [18].

Dynamic Time Warping (DTW) algorithm is one of the popular alignment strategies used for the alignment task in parallel VC [14, 73, 74]. The DTW algorithm tries to align two speech utterances globally, and not locally. To further improve the performance of DTW, phonetic information has been used to align two speech utterances locally (i.e, at the phoneme-level) [19]. However, the DTW algorithm assumes that the same phonemes spoken by the two speakers will be having similar features. However, spectral features are *not* speaker-independent. Hence, DTW will generate wrong aligned pairs. We termed these wrongly aligned feature pairs as outliers, since they were not following the overall trend of the data. In this thesis, we propose to analyze or identify these wrongly aligned pairs via novel outliers removal strategies, which are presented in Chapter 4 [5]. In addition, we also analyzed the impact of outliers removal on the quality of converted voices in this thesis.

Recently, more focus has been shifted towards the alignment step due to the need for building non-parallel VC systems for real-world applications [14]. Earlier, the idea of using a unit selection-based Text-To-Speech (TTS) synthesis system for generating parallel sentences from both the speakers was proposed, since durations can be mentioned to the TTS [75]. However, this approach needs more amount of training data to build high-quality TTS systems for both speakers. In addition, this approach needs text and hence, it is text-dependent. For text-independent and non-parallel VC systems, a unit selection approach is proposed, which finds best matching units from the target speaker's features based on the source speaker's features [76].

Later the idea of finding iteratively corresponding feature pairs based on Nearest Neighbor (NN) distance has become popular in the case of text-independent and non-parallel VC [6–8,77,78]. In particular, **I**terative combination of a **N**earest Neighbor search step and **C**onversion step **A**lignment (INCA) algorithm iteratively finds the corresponding feature pairs-based on the NN search, and then applies the conversion function until the convergence is achieved [6]. In the original paper of INCA, the convergence of the INCA algorithm was presented empirically [6]. In this thesis, we present a formal convergence theorem for the INCA algorithm [20]. To improve the performance of the INCA algorithm, recently, Temporal Context (TC) INCA algorithm was proposed [78]. To further improve the performance of the INCA, and TC-INCA algorithms, we also proposed to use dynamic features along with the static features for calculating NN pairs (as discussed in Chapter 5) [7].

The key issue of the NN-based alignment strategies is that they assume that perceptual distance correlates with the Euclidean distance [79]. However, it may not be true, and hence, the idea of metric learning is exploited for finding the NN pairs in the INCA algorithm in Chapter 5. We also proposed to use phonetic information in the NN-based alignment technique [9]. In particular, a novel Spectral Transition Measure (STM)-based alignment algorithm is proposed to estimate the phonetic boundaries [9,28], which is followed by the NN-based alignment in Chapter 5.

Recently, adaptation-based [80,81], and two-stage mapping techniques (that uses Phonetic PosteriorGram (PPG)) have been proposed to avoid the need of alignment in the VC tasks [11,31,82]. In this thesis, the issues associated with the PPG was identified in the coproposedntext of VC, and novel unsupervised Vocal Tract Length Normalized (VTLN) posteriorgram, and the Inter Mixture Weighted (IMW)-GMM posteriorgram have been proposed in the two-stage conversion techniques to avoid the need for alignment. In particular, the major contributions in this thesis are from the alignment perspectives are presented in detail in Chapter 4, Chapter 5, and Chapter 6 of this thesis. Next, we discuss the mapping step for the VC task.

2.5 Mapping Step for VC

Since the primary goal of the VC is to map the perceived speaker identity from a source speaker to a particular target speaker, the mapping techniques play a critical role in the design of the VC system. Most of the mapping techniques in VC literature are kind of supervised in nature for the VC task. Here, task is to find the mapping between source speaker's spectral features, i.e., $\mathbf{X}^{train} = [\mathbf{x}_1^{train}, \ldots, \mathbf{x}_N^{train}]$, and target speaker's spectral features, i.e., $\mathbf{Y}^{train} = [\mathbf{y}_1^{train}, \ldots, \mathbf{y}_N^{train}]$, which can be represented as $Y = \mathcal{F}(X)$. Here, $\mathbf{X}^{train}$, and $\mathbf{Y}^{train}$ are time-aligned, and N is the total number of corresponding frames obtained after the alignment. Once the training is done, at the time of conversion, source features, i.e., $\mathbf{X}^{test} = [\mathbf{x}_1^{test}, \ldots, \mathbf{x}_N^{test}]$ are passed through the conversion function, and the converted features, i.e., $\mathbf{Y}^{test} = [\mathbf{y}_1^{test}, \ldots, \mathbf{y}_N^{test}]$ are predicted. In the literature, the mapping function, i.e., $\mathcal{F}(.)$ maps the features frame-by-frame [14]. However, there have been attempts, which maps the frame along with its context than the frame-level mapping [14, 16].

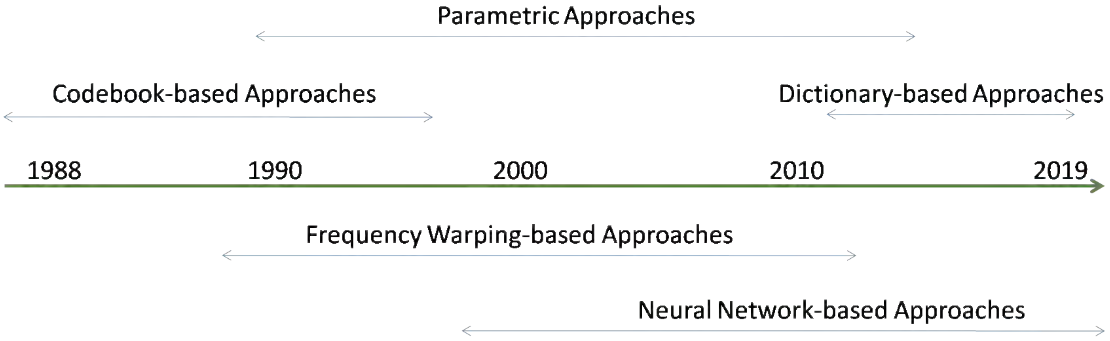


Figure 2.7: Selected chronological progress of different broad categories of mapping techniques for the VC task over the years. After [16].

Broadly, mapping techniques can be classified into five different categories, namely, codebook-based, parametric models-based, dictionary-based, frequency warping-based, and neural network-based approaches. These categories are described below.

- **Codebook-based mapping:** In order to reduce the number of source-target pairs in an optimized way, Vector Quantization (VQ) technique was proposed [74]. The key idea here is to generate M code vectors, i.e., $c_m^{\mathbf{x}}$, and $c_m^{\mathbf{y}}$ (where $m = [1, 2, \ldots, M]$), based on hard clustering using VQ technique on both the source and the target speaker's spectral features, respectively. Then, during conversion time, for a given source spectral features closest centroid vector from the source codebook is found, and the corresponding target codebook is predicted, i.e.,

$$\mathcal{F}_{VQ}(\mathbf{x}) = c_m^{\mathbf{y}}, \tag{2.1}$$

where m is the closest centroid vector, which is estimated for a given source spectral feature. Though the idea is simple here, it produces a discontinuous feature sequence. In order to reduce quantization error, the idea of fuzzy-VQ is also proposed, which primarily exploit the soft clustering [83]. This is achieved by assigning the continuous weights to each code vector, and at the time of conversion, converted features are predicted as a weighted sum of the code vectors. To further reduce quantization errors, and capture more variability, the idea of storing difference vector between source and the target code vector is stored, which is known as VQ-DIFF [84]. However, still, these approaches produce discontinuous converted feature sequence, which deteriorates the quality of the converted voices.

- **Parametric Approaches:** Linear Multivariate Regression (LMR) was one of the earliest parametric approaches that was applied for the VC task [56]. Here, the mapping function is calculated based on hard clustering of the source speaker space, which is given by,

$$\mathcal{F}_{LMR}(X) = A_m\mathbf{X} + b_m, \tag{2.2}$$

where A_m, and b_m are regression parameters of m^{th} code vector, which has minimum distance w.r.t. the given source spectral feature [56]. Similar to Fuzzy-VQ, this idea is also extended by assigning weights to each code vectors, i.e.,

$$\mathcal{F}_{wLMR} = \sum_{m=1}^{M} w_m^{\mathbf{x}}(A_m\mathbf{X} + b_m), \tag{2.3}$$

where $w_m^{\mathbf{x}}$ is the weight associated with each code vectors for a given speech frame. There are numerous attempts to estimate the regression parameters, among all the Joint Density Gaussian Mixture Model (JDGMM) was one of the popular approaches [19], which is discussed in detail in Chapter 3. In this method, source and target speakers' joint feature vectors are modeled using GMM, and the closed-form solution for the regression parameters is obtained, which is discussed in Chapter 3. Similar approach proposed in [85] apply GMM only to the source feature vector than the joint data and uses the least squares error optimization problem to obtain the regression parameters. The key issue with the GMM-based approach is that it involves computations of covariance matrices, which requires a huge amount of data, which is difficult to obtain in the context of real-world applications of VC, and hence, lead to *overfitting*. To address this issue, the idea of Partial Least Squares (PLS) regression is proposed [86]. PLS combines PCA and multivariate regression techniques to overcome the issues of overfitting. This approach is further extended using the dynamic kernels to capture nonlinear relationships [66]. Apart from these, there are also Radial Basis Function (RBF), and Support Vector Regressions (SVR)-based approaches have been proposed, which uses nonlinear kernels (such as Gaussian or polynomial) to transform source features in the high-dimension followed by the simple linear regression [87, 88]. Among all the parametric approaches, the Joint Density GMM-based approach, which uses Maximum Likelihood Estimation (MLE) criteria is considered as one of the best approaches. This method also used dynamic features and the Global Variance (GV) information to increase the variance of the generated features [16]. In this thesis, we discuss in detail the JDGMM-based method along with its key strengths and limitations in Chapter 3.

- **Frequency Warping (FW)-based Approaches:** Another important category of the mapping functions is the Frequency Warping (FW)-based approaches. In FW-based methods, the source spectrum is modified to match the frequency-axis of the target spectrum. The key motivation for these approaches is that both source and the target speaker will have different formant locations, formant bandwidth for the same utterances. The first attempt in the FW-based approach warped the source log spectrum-based on precomputed warping functions [56]. Later, most of the Vocal Tract Length Normalization (VTLN) techniques used in Automatic Speech Recognition (ASR) were applied to perform the VC task [89]. To further extend this approach, VTLN techniques were applied to multiple classes, and the iterative algorithm is proposed to estimate the VTLN parameters [42, 90]. Since FW-based methods do not remove any spectral details, it produces a high-quality voice after conversion. However, they do not modify the relative magnitude of the spectrum. Hence, the speaker similarity (SS) after conversion is not as successful as in the GMM-based VC systems. To overcome this problem, the FW-based method is complemented with Amplitude Scaling (AS) or residual spectrum compensation [42, 91, 92]. In this thesis work, we identify the key limitation of state-of-the-art AS technique, and proposed novel AS technique, which is discussed in Chapter 3.

- **Dictionary-based Approaches:** One of the most obvious approaches in the VC is to prepare a lookup table that has a source spectral feature as an input key and the target spectral feature as an output [14]. At the testing for a given source spectral feature, the nearest neighbor is identified from the lookup table and output is predicted. One of the key issues with these approaches is that the similarity of the source features does not necessarily mean similarity in the neighboring target features. Hence, it will cause a discontinuity in the generated parameter sequence [14]. To overcome this issue, Exemplar-based methods have been proposed, which assigns weights to all the target features [93, 94]. Since many frames will be assigned nonzero weights, it will results in oversmoothing due to averaging. To overcome this, Non-Negative Matrix Factorization (NMF) techniques have been proposed, which generates the sparse weights [93, 95].

- **Neural Network-based Approaches:** Recently, with the advent of deep learning, the neural network-based approaches have become very popular. Due to the nonlinear activation function associated with the neurons, the neural network-based approach is able to learn nonlinear relationships between the source and the target speakers' spectral representations. Earlier Artificial Neural Network (ANN) was used to map the formant frequency, and the spectral features using a simple feedforward network [96, 97]. Recently, with the availability of computational resources such as Graphical Processing Units (GPU)s, it has become possible to train neural networks with more number of hidden layers. ANN with more than two hidden layers is called Deep Neural Network (DNN) [98]. Due to more number of hidden layers, DNNs are known to capture more complex nonlinear relationships [98]. However, as the number of hidden layers increases, the parameters (i.e., weights, and biases of the network) to be estimated also increase. The random initialization of weights and biases of DNN results in a weak convergence, i.e., the likelihood gets stuck into the local minimum [98]. One of the possible solutions for the faster and better convergence for the training of DNN is to set initial parameters via pre-training of the network [98]. However, pre-training requires an extra network to train, which is time-consuming. In this thesis, we proposed the empirical analysis (which is discussed in detail in Chapter 3) to show that the need for pre-training can be avoided in the context of VC by using advanced deep learning strategies [4]. Recently, many DNN architectures have been successfully applied in VC tasks. For example, Restricted Boltzmann Machine (RBM) [99], Recurrent Neural Network (RNN)-based architecture [100], Long Short Term Memory (LSTM) [82, 101, 102], sequence-to-sequence [103, 104], Variational AutoEncoder (VAE) [105,106], Generative Adversarial Network [10,11,103,107,108], Cycle-consistent Network [109], etc. In this thesis work, we studied recent GAN-based architectures in detail and proposed novel deep learning architectures from the mapping perspectives in Chapter 3.

Recently, the mapping function stage has gained a lot of interest with the advent of deep learning. Details of the mapping techniques along with their strengths and limitations are presented in Chapter 3. This thesis also contributes significantly from the mapping perspectives, which are discussed in Chapter 3.

2.6 Prosody Conversion for VC

Even though, we cannot define speech prosody explicitly, there are few attempts such as pitch, intonation, rhythm, duration, loudness, etc., which are counterparts to the speech prosody [110, 111]. It has been shown that when professional mimicry artists (i.e., voice imitators) try to mimic a particular target speaker, he or she only adjusts the average statistics of the F_0 contour rather than completely mimicking the F_0 contour of the target speaker [36, 37]. Hence, most of the VC techniques in the literature are more focused on mapping spectral features than the prosodic features [14]. In this work, F_0 is converted via simple linear transformation using the global statistics of the F_0 parameters. In particular, F_0 is transformed by [16, 112, 113], i.e.,

$$\log(F_0^y) = \frac{\sigma^y}{\sigma^x}(\log(F_0^x) - \mu^x) + \mu^y, \tag{2.4}$$

where μ^x, and μ^y are the mean of the F_0 contour from the source and target speakers, respectively. Similarly, σ^x, and σ^y are the standard deviations of the F_0 contour obtained from the source and the target speaker, respectively. Though prosodic features are important for speaker identity, developing sophisticated prosody models involve great challenges (which are briefly described below), and it continues to be an open research problem. However, it has been shown recently that by mapping F_0 contour more closely to the target speaker's F_0 pattern will result in better speaker similarity [114–117]. Some of the technological challenges in incorporating speech prosody for VC tasks are as follows.

- The prominence (stress) given at a particular word in an utterance also results into its effect onto the neighboring words (in terms of their duration of sound units, a shift in spectrum energy densities at the higher frequency range, rise in F_0, etc.) Modeling these effects for realistic applications, such as VC requires an understanding of more broad contextual interactions, which is phenomenally complex, and highly depends on *cognitive* factors, which are in fact very subjective (i.e., different for source and the target speakers).

- With respect to the above-mentioned point, the subjective or cognitive factors will then impose additional challenges in creating differences in the duration of various sound units for the alignment task and also for the mapping task to a certain extent!

2.7 Databases and Evaluations of the Converted Voices

2.7.1 VC Databases

Till 2016, no standard database was available for the VC task. With Voice Conversion Challenge (VCC) 2016 and 2018, standard corpora have been designed for the VC task. Earlier, research papers reported results on the TIMIT, CMU-ARCTIC, MOCHA-TIMIT, and VOICES corpora. Table 2.1 presents the statistics of the databases used in this thesis.

Table 2.1: Statistics of databases used in VC Systems

Database		No. of Speakers		No. of Utterances	
		Male	Female	Training	Testing
CMU-ARCTIC [18]	Source	1	1	1078	54
	Target	1	1	1078	54
VCC 2016 [118]	Source	2	3	162	54
	Target	3	2	162	54
VCC 2018 [119] HUB Task	Source	2	2	81	35
	Target	2	2	81	35
VCC 2018 [119] SPOKE Task	Source	2	2	81	35
	Target	2	2	81	35

2.7.2 Evaluations of Converted Voices

Converted voices are evaluated using primarily via objective, and subjective measures. One of the standard objective measures in the VC is the Mel Cepstral Distortion (MCD). This objective measure needs parallel utterances from the source and the target speakers. The length of the converted voices, and the target will not be the same. Hence, first it is aligned via the DTW algorithm. Then, differences between converted, and the target cepstral features are calculated. The traditional objective measure, MCD is considered here, which is given by [16]:

$$MCD \text{ [dB]} = \frac{10}{ln10}\sqrt{2\sum_{i=2}^{d}(x_i^t - x_i^c)^2} \ , \tag{2.5}$$

where x_i^t and x_i^c are the i^{th} coefficient of MCC features corresponding to the target, and the converted voice, respectively. In addition, d is the dimension of the MCC feature vector. The smaller value of the MCD means that the VC system is performing relatively better compared to the VC system that is having a higher value of MCD. It has been well-known that objective measures do not correlate well with the subjective measures (i.e., human perception due to its high level of sophistication) [12, 14]. Similar challenges are associated with the objective evaluations of TTS voice [120, 121]. Hence, designing novel objective measure, which correlates well with the subjective scores continues to be an open research problem. In this thesis, we propose to use articulatory features obtained via Acoustic-to-Articulatory Inversion (AAI) technique for the objective evaluation [12]. Details of the proposed approach are presented in Chapter 7. In addition, recently the idea of using ASR, and the Spoof Speech Detection (SSD) system for the quality assessment of the converted voices have been proposed [119, 122].

Subjective evaluations are necessary to evaluate the performance of the VC system. In this task, human listeners participate in order to assess the VC system. There are primarily three aspects that are evaluated during the subjective test in the context of the VC task, namely, speaker similarity, speech quality, speech intelligibility, which can be explained as below.

- **Speaker Similarity:** Since the task is to convert perceived speaker identity in the VC, subjects are asked to rate converted voices based on similarity w.r.t. the target speaker.

- **Speech Quality:** Here, subjects are asked to rate the converted voices based on audible artifacts and naturalness.

- **Speech Intelligibility:** As the task of the VC is to convert perceived speaker identity without changing the linguistic content of the speech, assessment of converted voice based on speech intelligibility is also an important factor. In particular, human listeners have to assess the converted voices based on the clarity of the linguistic message present in it.

Mean Opinion Score (MOS), and ABX tests have been conducted in this thesis. In MOS tests, subjects have been asked to rate the success w.r.t. the above-mentioned aspect of the converted voices on a 5-point scale (5-Excellent, 4-Good, 3-Average, 2-poor, and 1-Bad). On the other hand, ABX tests have been conducted to measure the relative comparison of the above-mentioned aspects of the converted voices obtained from the two different VC systems. Details of the subjective tests are also presented throughout the thesis at various locations. High-quality headphones have been used for subjective evaluations. All the subjective tests have been conducted in a quiet environment. All the subjective results reported in the thesis have been taken from a statistically meaningful number of subjects in the thesis. Subjects were coming from the age group of 18-30 years, and they did not have any known hearing impairments.

2.8 Chapter Summary

In this chapter, we presented an overview of the VC system with special emphasis on alignment and mappings. In particular, different stages of VC system building have been discussed in detail along with the selected chronological progress. Furthermore, details of alignment and mapping steps were presented to identify the gap area in the VC field and hence, locate the problem of research for this thesis. In the next chapter, we will discuss in detail the contributions of the thesis from the mapping perspective for the VC task.

CHAPTER 3
Mapping Techniques

3.1 Introduction

In the last chapter, literature search on various issues in VC was presented. Among different stages of the VC system buildings, we focus on the mapping step in this chapter. Since the primary goal of the VC is to map the perceived speaker identity from a source speaker to a particular target speaker, the mapping techniques play a critical role for the design of VC system. Most of the mapping techniques in VC literature are kind of supervised in nature for the VC task. These techniques require corresponding feature pairs from the source and the target speakers. Obtaining such a pair in the context of parallel and non-parallel training cases is a difficult research problem, which is discussed in Chapter 4 and Chapter 5, respectively. However, we will assume that we have access to the corresponding feature pairs in this chapter. The progress made in this thesis from mapping perspective in the context of VC task is discussed in this chapter.

3.2 Joint Density Gaussian Mixture Model (JDGMM)

Among all the available VC techniques, GMM-based VC technique is considered as one of the state-of-the-art mapping technique [14]. Joint Density (JD) GMM-based VC technique finds the mapping function between the source and target speakers' feature vectors. Let the $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$, and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_M]$ be the spectral feature vectors corresponding to the source and the target speakers, respectively. Here, $\mathbf{x}_n \in \mathbb{R}^d$, and $\mathbf{y}_n \in \mathbb{R}^d$ (i.e., $\mathbf{x}$ and $\mathbf{y}$ are d-dimensional feature vectors). The joint vector, $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_r, ..., \mathbf{z}_K]$, is formed after aligning the spectral features, $\mathbf{x}_n$ and $\mathbf{y}_n$ using the Dynamic Time Warping (DTW) algorithm, where $\mathbf{z}_r = [\mathbf{x}_n^T, \mathbf{y}_m^T]^T \in \mathbb{R}^{2d}$. Furthermore, the joint vector is modeled by the GMM as follows [19]:

$$P(\mathbf{Z}) = \sum_{m=1}^{N_C} \omega_m^{(\mathbf{z})} \mathcal{N}(\mathbf{Z}|\mu_m^{(\mathbf{z})}, \Sigma_m^{(\mathbf{z})}), \tag{3.1}$$

where $\mu_m^{(\mathbf{z})} = \begin{bmatrix} \mu_m^{(\mathbf{x})} \\ \mu_m^{(\mathbf{y})} \end{bmatrix}$ and $\Sigma_m^{(\mathbf{z})} = \begin{bmatrix} \Sigma_m^{(\mathbf{xx})} & \Sigma_m^{(\mathbf{xy})} \\ \Sigma_m^{(\mathbf{yx})} & \Sigma_m^{(\mathbf{yy})} \end{bmatrix}$ are the mean vector, and covariance matrix of the m^{th} mixture component, respectively, N_C total number of mixtures, and ω_m is the weight associated with the m^{th} mixture component with the constraint $\sum_{m=1}^{N_C} \omega_m^{(\mathbf{z})} = 1$ (for total probability). During the training of GMM, the model parameters, $\lambda^{(\mathbf{z})} = \{w_m^{(\mathbf{z})}, \mu_m^{(\mathbf{z})}, \Sigma_m^{(\mathbf{z})}\}$ are estimated using the Expectation-Maximization (EM) algorithm [123]. Once the training is completed, the *mapping* function (i.e., $\mathcal{F}(\cdot)$) is learned either using the Minimum Mean Squared Error (MMSE) criteria [19] or using the Maximum Likelihood Estimation (MLE)-based criteria [16]. Since the conditional expectation is the best MMSE estimator (the proof of this result is discussed in Apeendix A), the mapping function $\mathcal{F}(\cdot)$ is given by [19]:

$$\begin{aligned}
\hat{\mathbf{y}}_t &= \mathcal{F}(\mathbf{x}_t), \\
&= E[\mathbf{y}_t|\mathbf{x}_t], \\
&= \int P(\mathbf{y}_t|\mathbf{x}_t, \lambda)\mathbf{y}_t d\mathbf{y}_t, \\
&= \int \sum_{m=1}^{N_c} P(m|\mathbf{x}_t, \lambda^z) P(\mathbf{y}_t|\mathbf{x}_t, m, \lambda^{\mathbf{z}})\mathbf{y}_t d\mathbf{y}_t, \\
&= \sum_{m=1}^{N_c} P(m|\mathbf{x}_t, \lambda^{\mathbf{z}})(E_{mt}^{(\mathbf{y})}),
\end{aligned}$$

$$\hat{\mathbf{y}}_t = \sum_{m=1}^{N_c} P(m|\mathbf{x}_t, \lambda^{\mathbf{z}})(\mu_m^{(\mathbf{y})} + \Sigma_m^{(\mathbf{yx})}(\Sigma_m^{(\mathbf{xx})})^{-1}(\mathbf{x}_t - \mu_m^{(\mathbf{x})})), \tag{3.2}$$

where $P(m|\mathbf{x}_t, \lambda^{\mathbf{z}}) = \frac{\omega_m \mathcal{N}(\mathbf{x}|\mu_m^{\mathbf{x}}, \Sigma_m^{\mathbf{xx}}))}{\sum_{k=1}^{M} \omega_k \mathcal{N}(\mathbf{x}|\mu_k^{\mathbf{x}}, \Sigma_k^{\mathbf{xx}}))}$ is the posterior probability of the source vector, $\mathbf{x}_t$, for the m^{th} Gaussian component. In addition, $E_{mt}^{(\mathbf{y})}$ is the mean of the conditional distribution $P(\mathbf{y}_t|\mathbf{x}_t, \lambda^{\mathbf{z}})$, which is derived in Appendix B. One of the interesting points about Gaussian is that, whenever two variables are jointly Gaussian its conditional distribution is also Gaussian. Derivation of closed form solution for $E_{mt}^{(\mathbf{y})} = \mu_m^{(\mathbf{y})} + \Sigma_m^{(\mathbf{yx})}(\Sigma_m^{(\mathbf{xx})})^{-1}(\mathbf{x}_t - \mu_m^{(\mathbf{x})})$ is given in the Appendix B. Statistical averaging in eq. (3.2) leads to *oversmoothing* of the converted voice. The *oversmoothing* is undesirable, as it deteriorates the quality of the converted voice. To alleviate this, the use of dynamic features, and Global Variance (GV) enhancement methods were proposed in [16]. Schematic representation of conditional probability density function (*pdf*) in the JDGMM is shown in Figure 3.1.

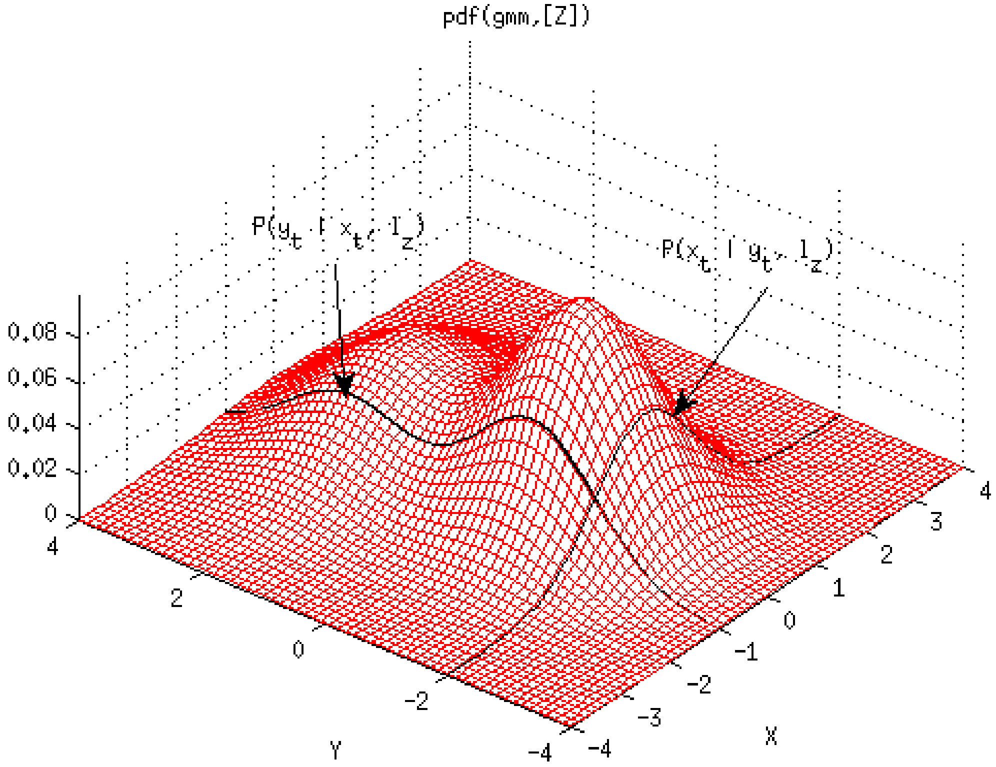


Figure 3.1: Schematic representation of conditional *pdf* in the JDGMM. After [19,20].

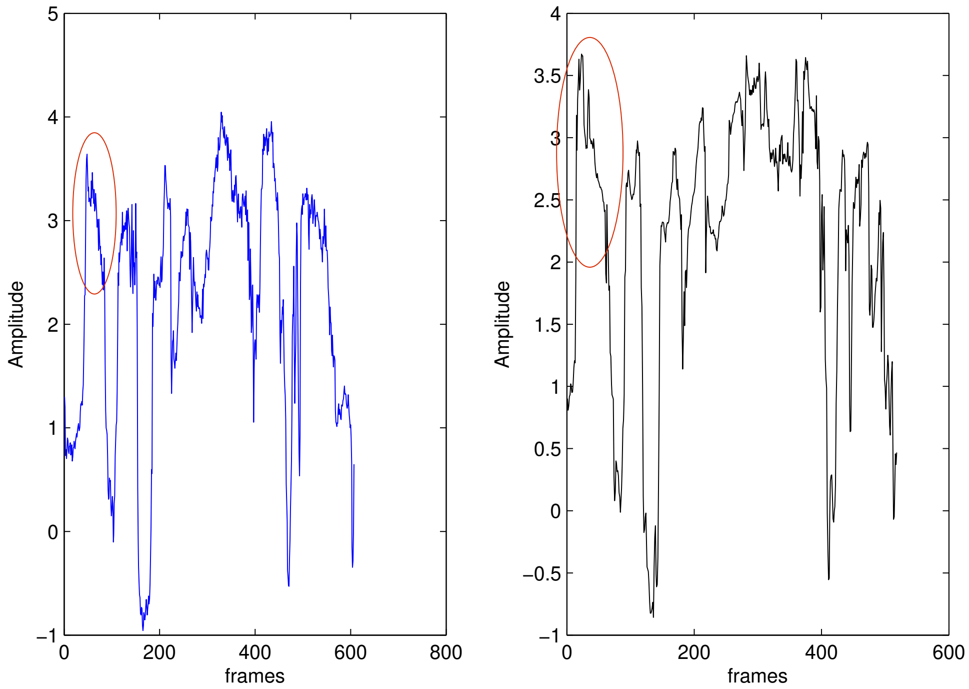


Figure 3.2: 2^{nd} Mel Cepstral Coefficient for (a) actual target speaker, and (b) JDGMM-based converted voice.

3.3 Bilinear Frequency Warping (BLFW) with Proposed Amplitude Scaling

The GMM-based VC method transforms the overall gross spectral characteristics very well. However, the finer details are *not* well transformed due to the oversmoothing (as shown in Figure 3.2). Apart from the use of dynamic features and GV, exemplar-based non-parametric techniques were also proposed which directly uses the target speech exemplars to synthesize the converted speech and hence, keep more spectral details [93–95, 124, 125].

Apart from this, there are Frequency Warping (FW)-based methods in which the source spectrum is modified to match the frequency-axis of the target spectrum (as shown in Figure 3.3). Among various FW-based methods [21, 42, 56, 126–129], BLFW method has been selected. As discussed in [42], the BLFW-based VC can be formulated in the parametric-domain. In addition, the BLFW do not show a locally irregular behavior compared to the piecewise learning-based FW methods. Furthermore, the number of parameters to be learnt is smaller which makes it suitable in the context of *overfitting* [42].

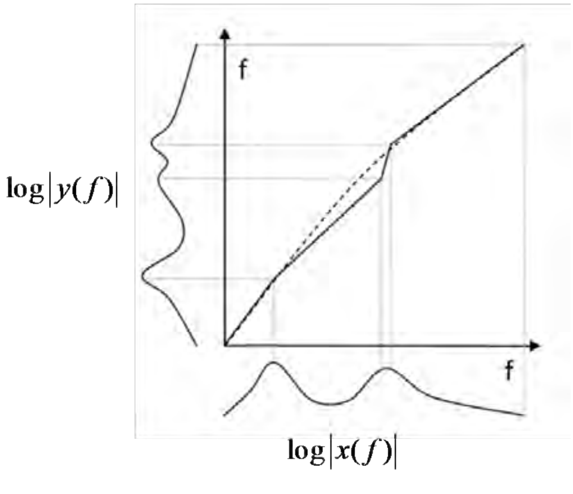


Figure 3.3: Basic idea of frequency warping-based VC. After [21].

Since FW-based methods do not remove any spectral details, it produces a high quality voice after conversion. However, they do not modify the relative magnitude of the spectrum. Hence, the speaker similarity (SS) after conversion is not as successful as in the GMM-based VC systems. To overcome this problem, FW-based method is complemented with Amplitude Scaling (AS) or residual spectrum compensation [42, 91, 92]. The AS modifies the vertical-axis of the warped spectrum. The AS operation in the state-of-the-art BLFW+AS method assumes the perfect match between the warped, and the target formant structures which is not possible in practice [42]. As a result, the AS vector not only contains information related to the amplitude of the spectrum but also some information related the location of the formant frequencies (which will add spurious peaks in the warped spectrum). Hence, the quality of a converted voice is expected to be degraded.

To eliminate such spurious peaks, we propose a novel AS technique at spectrum-level. The proposed AS transfers the spectral range of the warped-only spectrum to the spectral range of GMM-based spectrum. Several attempts have been made to combine the two state-of-the-art methods, namely, GMM and FW-based methods, in order to exploit the advantages of both the methods [91], [130], [131]. Similarly, our proposed AS method also combines the knowledge of this two state-of-the-art methods to obtain a better quality compared to the BLFW+AS method.

3.3.1 BLFW-based VC

For a given d-dimensional feature vector $\mathbf{x}$, its frequency-warped feature vector $\mathbf{y}$ is given by

$$\mathbf{y} = W_\alpha \mathbf{x}, \tag{3.3}$$

$$W_\alpha = \begin{bmatrix} 1-\alpha^2 & 2\alpha - 2\alpha^3 & \dots \\ -\alpha + \alpha^3 & 1 - 4\alpha^2 + 3\alpha^4 & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}, \tag{3.4}$$

where W_α (called a warping matrix) has been expressed without considering the 0^{th} cepstral coefficient. The BLFW method uses an allpass transform that is given by [42]:

$$Q(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \tag{3.5}$$

where $|\alpha| < 1$ and if $z = e^{j\omega}$, the frequency response of the allpass transform can be given by:

$$Q(e^{j\omega}) = \frac{e^{-j\omega} - \alpha}{1 - \alpha e^{-j\omega}}. \tag{3.6}$$

The magnitude response of the allpass filter is shown in Figure 3.4. It can be clearly seen that it will allow all the frequencies to pass. The relation between the warped frequency, and the original frequency is given by:

$$\omega_\alpha = tan^{-1}\left[\frac{(1-\alpha^2)sin\omega}{(1+\alpha^2)cos\omega - 2\alpha}\right]. \tag{3.7}$$

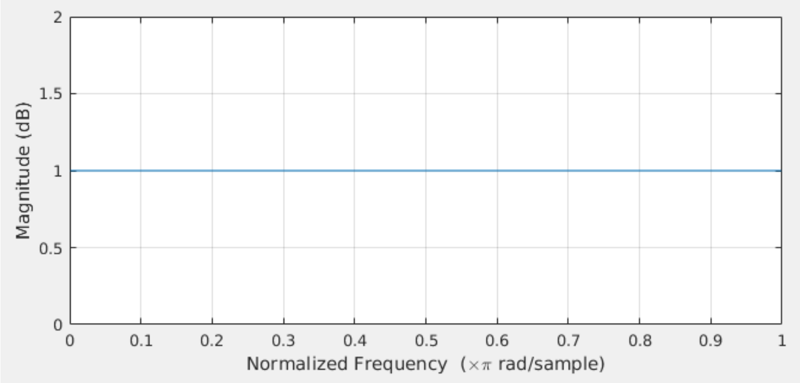


Figure 3.4: Frequency response of the allpass transform.

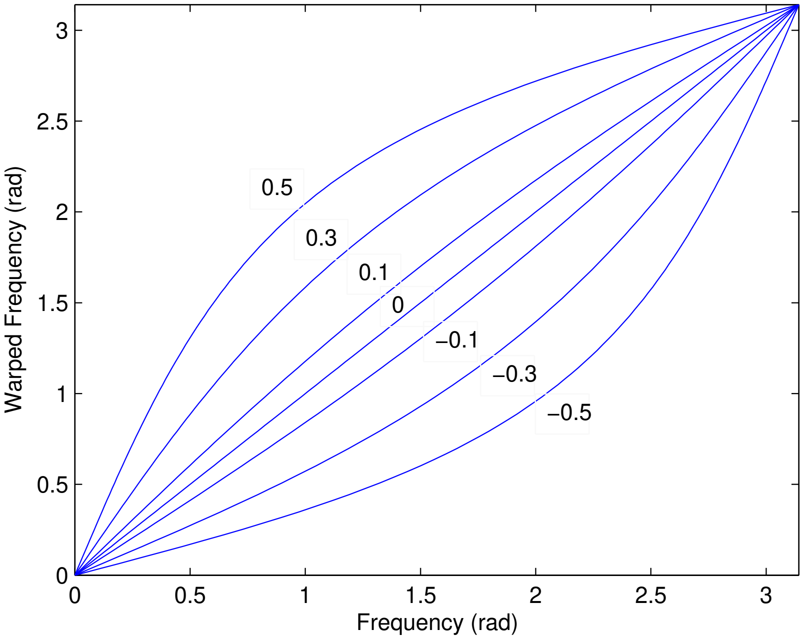


Figure 3.5: Shape of a BLFW function for different values of α. After [22].

Figure 3.5 shows the shape of this curve for different values of α estimated from eq. (3.7). From Figure 3.5, it can be observed that the positive values of α move the warped frequencies (i.e., possible formant) to the higher frequencies (as in case of a male-to-female conversion), and similarly, negative values of α move the formants to the lower frequencies (as in case of a female-to-male conversion). Thus, it maintains the *inverse* relationship between the vocal tract length, and the formant frequencies, which is given by [17]:

$$f_k = \frac{(2k+1).c}{4l}, \tag{3.8}$$

where f_k is the k^{th} formant frequency, c is the speed of sound (i.e., 350 m/s), and l

is the length of the vocal tract system.

Instead of finding global warping factor for entire training database, here, we propose to estimate warping factor for each component of GMM. GMM is modeled on training database of a source speaker (i.e., λ). FW factor α_m, and AS vector s_m are associated with each components of GMM and hence, the conversion function is given by [42]:

$$\mathbf{y} = W_{\alpha(\mathbf{x},\lambda)}\mathbf{x} + s(\mathbf{x}, \lambda), \tag{3.9}$$

where $\alpha(x, \lambda)$ and $s(x, \lambda)$ are the result of combining the basis warping factors, and the AS vectors of all the components of λ, respectively, which is given by:

$$\alpha(\mathbf{x}, \lambda) = \sum_{m=1}^{N_c} p_m^{(\lambda)}(x)\alpha_m, \quad s(\mathbf{x}, \lambda) = \sum_{m=1}^{N_c} p_m^{(\lambda)}(\mathbf{x})s_m, \tag{3.10}$$

where $p_m^\lambda(\mathbf{x})$ is the probability that $\mathbf{x}$ belongs to m^{th} mixture component of λ, and N_c is the total number of mixture component. Given the aligned source and target feature vectors and GMM trained on the source speaker data, i.e., λ, the warping factor α_m is first estimated by minimizing the error of warping only conversion, which is given by:

$$\epsilon^{(\alpha)} = \sum_{n=1}^{N} ||\mathbf{y}_n - W_{\alpha(\mathbf{x}_n,\lambda)}\mathbf{x}_n||^2. \tag{3.11}$$

The iterative procedure proposed in [42] for calculating a set of $\{\alpha_m\}$ for minimizing the eq. (3.9) is used here.

3.3.2 State-of-the-art Amplitude Scaling (AS) Method

Once $\{\alpha_m\}$'s are estimated, the $\{s_m\}$ that minimizes the error between the warped, and target vectors is given by [42]:

$$\epsilon^{(s)} = \sum_{n=1}^{N} ||\mathbf{r}_n - s(\mathbf{x}_n, \lambda)||^2, \tag{3.12}$$

where $\mathbf{r}_n = \mathbf{y}_n - W_\alpha(\mathbf{x}_n)$. This means that calculating the least square solutions of system, i.e., $P \cdot S = R$, where

$$P_{N\times m} = \begin{bmatrix} p_1^{(\lambda)}(\mathbf{x}_1) & \dots & p_{N_c}^{(\lambda)}(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ p_1^{(\lambda)}(\mathbf{x}_N) & \dots & p_{N_c}^{(\lambda)}(\mathbf{x}_N) \end{bmatrix}, \tag{3.13}$$

$$\text{and } S_{m\times 1} = [s_1 \dots s_m]^T, \quad R_{N\times 1} = [r_1 \dots r_N]^T. \tag{3.14}$$

The least square solution via l^2 norm minimization is given by:

$$S_{opt} = (P^T P)^{-1} P^T R. \tag{3.15}$$

The AS vector should compensate for the different formant amplitudes. In some of the cases where the warped formants do not coincide with the actual target formants, AS vector is expected to capture the mixed information about the intensity as well as the location of the formant which is potentially harmful to the voice quality of a converted voice [22].

3.3.3 Proposed AS Method

The AS operation in the above mentioned method assumes that there will be a perfect match between warped, and target formant structures which is not possible in practice. Hence, the AS operation will induce spurious peaks, giving the perceptual impression of a wrong formant locations leading to a deterioration of speech quality in the converted speech signal. It can be seen from Figure 3.6 that BLFW+AS method adds spurious peaks in only BLFW warped spectrum (OBLFW). Essentially, AS operation should alter only the amplitudes of the warped spectrum (i.e., intensity of the formant) which is not the case. Therefore, we propose the following linear transformation at the spectrum-level:

$$\hat{y}_t(e^{j\omega}) = \frac{(b_3 - b_4)}{(b_1 - b_2)}(\hat{x}_t(e^{j\omega}) - b_2) + b_4, \tag{3.16}$$

where $\hat{x}_t(e^{j\omega})$ is the warped-only spectrum, and

$$\begin{aligned} b_1 &= max(\hat{x}_t(e^{j\omega})), \quad b_2 = min(\hat{x}_t(e^{j\omega})), \\ b_3 &= max(\hat{x}_{tgmm}(e^{j\omega})), b_4 = min(\hat{x}_{tgmm}(e^{j\omega})), \end{aligned} \tag{3.17}$$

where $max()$ and $min()$ will find the maximum and minimum value of a spectrum, respectively. In addition, $\hat{x}_{tgmm}(e^{j\omega})$ is the converted spectrum using JDGMM method.

Here, the proposed AS method transforms the spectral range of OBLFW spectrum to the spectral range of GMM-based converted spectrum. Since GMM-based VC transfers well the gross spectral characteristics, spectral range of converted spectrum obtained using GMM-based VC will be helpful to compensate the amplitude difference between warping-based spectrum, and the true target spectrum. As the proposed method uses the spectral range information instead of a finer details of GMM-based converted spectrum, it is free from the issue of *over-smoothing*. It can be seen from Figure 3.6 that the proposed AS (i.e., BLFW+PAS) will not add any spurious peaks and will compensate only for the amplitude difference without affecting the quality of a converted speech. Here, we would like to show the effectiveness of proposed AS method over the state-of-the-art AS on the BLFW-based warped spectrum. Hence, the GMM-based spectrum and the actual target spectrum is not shown in Figure 3.6. Similar spurious peaks are observed for the state-of-the-art AS methods for most of the frames.

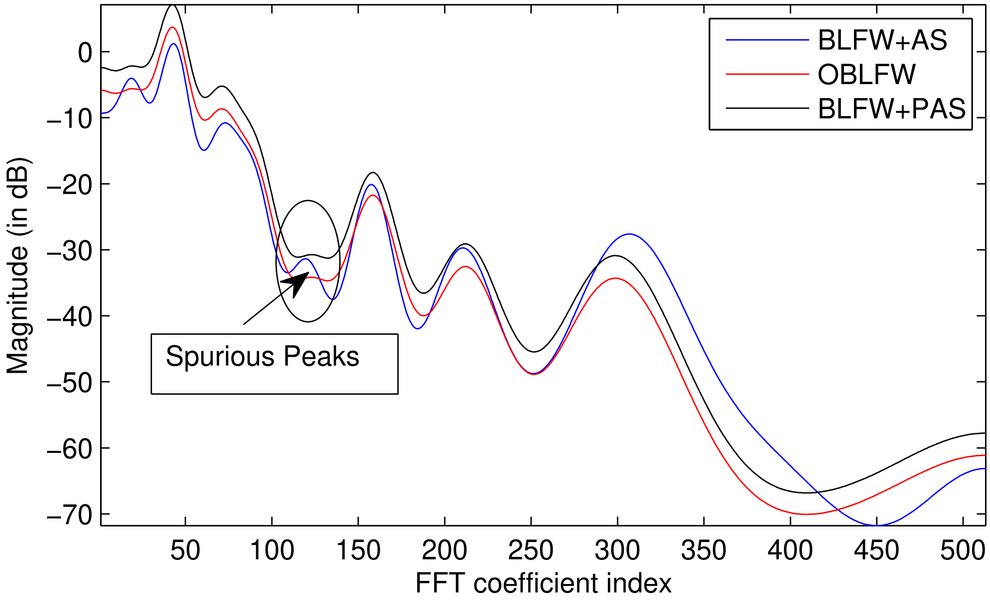


Figure 3.6: Converted spectrum using various VC methods. After [22].

3.3.4 Experimental Results

First VC challenge database is used in this work [118]. We have built in total *25* systems for each source-target speaker-pair using JDGMM-based method, BLFW+AS method, and the proposed method (i.e., BLFW+PAS). *25*-D Mel cepstral coefficients (MCEPs) (including the 0^{th} coefficient), and *1*-D F_0 per frame (with *25* ms frame duration, and *5* ms frame shift) have been used. The Dynamic Time Warping (DTW) algorithm has been used to align parallel training corpora [73]. For JDGMM-based system, and for training of source GMM in the case of BLFW, we have taken different values of number of mixture components. For example, *m=16, 32, 64, 128*, and selected the one which leads to the optimum MCD. We used a mean-variance (MV) transform method for F_0 transformation. AHOCODER has been used for the analysis-synthesis framework [67].

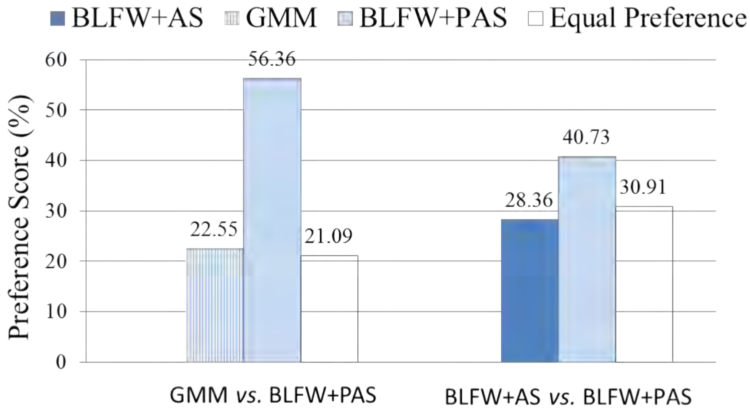


Figure 3.7: XAB test analysis for voice quality along with 95 % confidence interval (margin of error: 0.048 for GMM *vs.* BLFW+PAS, and 0.05 for BLFW+AS *vs.* BLFW+PAS). After [22].

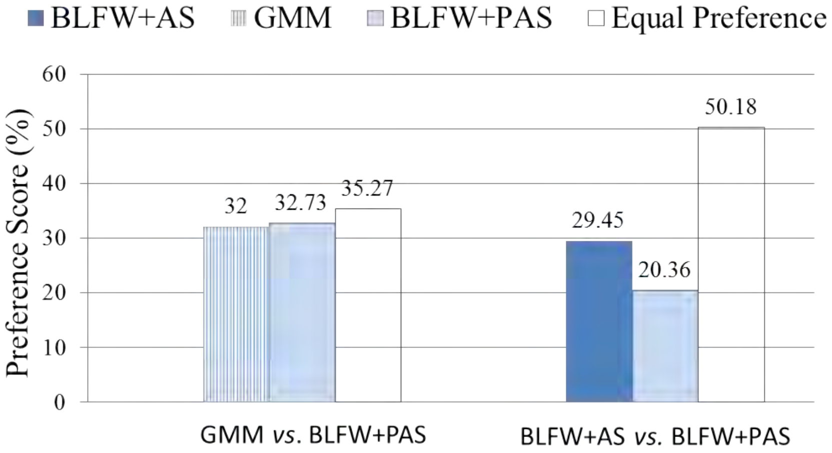


Figure 3.8: XAB test analysis for speaker similarity along with 95 % confidence interval (margin of error: 0.05 for both the cases). After [22].

For the subjective evaluation, comparative subjective test, namely, XAB test has been selected. Subjects were asked to prefer from the randomly played *A* and *B* samples (generated from two different approaches) which is having better voice quality, and *speaker similarity* (SS) with reference to the actual target sample X. In addition, the subjects can select equal preference in the case of samples that are *perceptually* similar. XAB test was performed separately between JDGMM, and BLFW+PAS (i.e., the proposed method) and between BLFW+AS, and BLFW+PAS. Figure 3.7 and Figure 3.8 show the MOS obtained from the *15* subjects (*5* females and *10* males with no known hearing impairments and age varies between 18 to 30 years) from total *375* samples for voice quality and SS, respectively. It is clear from the results that in terms of voice quality, the proposed AS system is preferred

56.36 % times, whereas the GMM-based system is preferred *22.55* % times by the subjects. Similarly, BLFW+PAS is preferred *40.73* % times whereas BLFW+AS is preferred *28.36* % times by the subjects.

The speaker identity conversion was *0.73* % times more preferred over GMM-based method. Though the proposed system *9.09* % times less preferred over BLFW+AS system, *50.18* % times subjects have given an equal preference to the proposed system and BLFW+AS. The less preference for speaker similarity of the proposed system compared to the state-of-the-art BLFW+AS clearly indicates that actual shape of the spectral trajectory also matters for better speaker identity conversion in addition to the formant locations and its amplitude [52]. However, modifying the spectral details will affect the voice quality. Hence, there is a quality conversion trade-offs. Similar trade-offs were observed by the other studies in the literature [92, 127, 132].

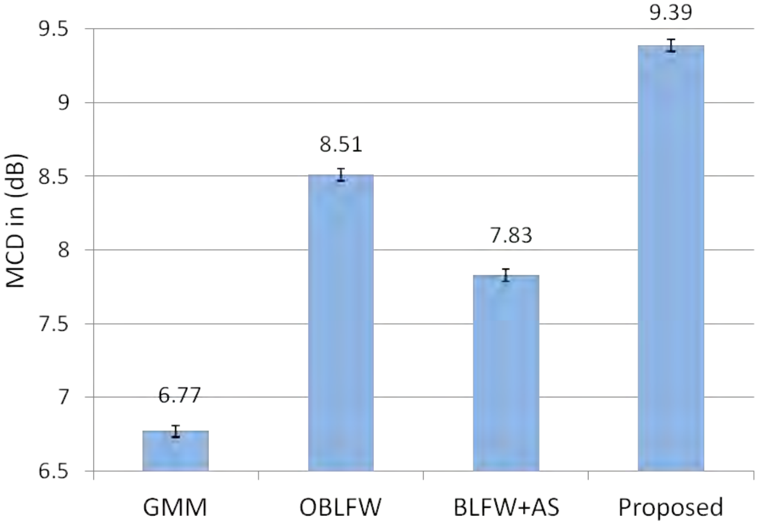


Figure 3.9: MCD analysis for various systems along with *95* % confidence interval (margin of error: 0.04 for all the systems). After [22].

For objective evaluation, the traditional MCD is used. It can be seen from Figure 3.9 that the proposed method gives the higher MCD values compared to the GMM-based VC. The BLFW method moves the formants towards their image in the target speaker's spectrum. Thereafter, the proposed AS will modify the amplitude of warped spectrum instead of matching the actual target spectral details. Hence, it will not get less MCD scores compared to the GMM-based VC (as shown in Figure 3.9). In addition, it has been observed in the literature that MCD does *not* correlate well with subjective score for FW-based VC [12,42,92,127]. However, MCD is used here for comparing the relative performance of same type of VC for selecting optimum number of mixture components.

Apart from these signal processing-based methods, neural network-based methods have also become very popular recently [14]. Next Section briefly summarizes the neural network-based approaches to estimate the mapping function.

3.4 Neural Networks-based Methods

3.4.1 Artificial Neural Network (ANN)

In the ANN-based VC, the relationship between the spectral feature vectors $\mathbf{X}$, and $\mathbf{Y}$ is obtained using the ANN that has L hidden layers. The first, last, and the middle layers of the ANN are called as an input, output, and the hidden layer, respectively. Here, each layer performs either nonlinear or linear transformation. The transformation at the i^{th} layer is given by [98]:

$$\mathbf{h}_{i+1} = f(\mathbf{W}_i^T \mathbf{h}_i + \mathbf{b}_i), \tag{3.18}$$

where $\mathbf{h}_i$, $\mathbf{h}_{i+1}$, $\mathbf{W}_i$, $\mathbf{b}_i$ are called as the input, output, weights, and the bias of i^{th} layer, and f is an activation function, which is generally nonlinear (such as tangent hyperbolic, sigmoid, Rectifier Linear Unit (ReLU), etc.) or linear. $\mathbf{h}_1 = \mathbf{X}$, and $\mathbf{h}_{L+1} = \mathbf{Y}$ are the input, and output layers of the ANN, respectively. L is the total number of hidden layers and hence, $L+1$ is the output layer. The Stochastic Gradient Descent (SGD) algorithm is used to train the weights and biases of the ANN such that MSE, i.e., $E = ||\mathbf{Y} - \hat{\mathbf{Y}}||^2$ is minimized [98]. Here, $\hat{\mathbf{Y}}$ is the predicted output from the network, and $\mathbf{Y}$ is the true output.

3.4.2 Deep Neural Network (DNN)

The ANN with more than two hidden layers (i.e., $L > 2$) is called as DNN [98]. The DNN can capture the more complex relationships between the source and the target speakers' spectral features due to more number of hidden layers [47]. However, as the number of hidden layers increases, the parameters (i.e., weights, and biases of the network) to be estimated also increase. The random initialization of weights and the biases of DNN results in a weak convergence, i.e., the likelihood gets stuck into the local minimum [98]. One of the possible solutions for the faster and better convergence for the training of DNN is to set initial parameters via pre-training of the network [98]. In particular, we have used a stacked Denoising AutoEncoder (DAE) that is created by stacking the layers of the AutoEncoders (AE) for the pre-training of DNNs [98].

3.5 Strategies for Training DNN for VC

3.5.1 Pre-training of DNN

The random initialization of weights and the biases of the DNN results in poor convergence, i.e., the likelihood will get stuck into local minima [133]. One of the possible solutions for faster and better convergence for the training of DNN is to set initial parameters via pre-training of the network [134]. In particular, we have used a DAE that is created by stacking of layers of the autoencoders for pre-training of DNN [135]. The baseline DNN consists of encoding layers of DAE, followed by shallow ANN, and decoding layers of DAE [47].

3.5.2 Whether To Pretrain DNN or Not?: An Empirical Analysis

DNN-based VC techniques suffer from the issue of overfitting due to less amount of available training data from the target speaker. To alleviate this, pre-training is used for better initialization of the DNN parameters, which leads to faster convergence of parameters. Greedy layerwise pre-training of the stacked Restricted Boltzmann Machine (RBM) or the stacked De-noising AutoEncoder (DAE) is used with extra available speaker-pairs' data. This pre-training is time-consuming, and requires a separate network to learn the parameters of the network. In this work, we propose to analyze the DNN training strategies for the VC task, specifically with and without pre-training. In particular, we investigate whether an extra pre-training step could be avoided by using recent advances in the deep learning literature.

3.5.3 Regularization

Sometimes during the DNN training, weights of the neighboring neurons become more dependent on the current neuron's weight, and this dependency is called complex co-adaptation [136]. Hence, if the neurons are randomly dropped out, the neighboring neurons will have to step in and make accurate predictions for the missing neurons, which will make our network to generalize itself very well, and also make it less sensitive to the overfitting. Dropout is one of the most simple yet very effective ways of preventing overfitting in DNN [136]. We used dropout as a regularization in our DNN training. Here, dropout is not applied at the input and the output layers. The term dropout refers to randomly dropping out neurons with probability, p. Applying a dropout to DNN can be considered as a multiplying neural network activation with a binary mask (also known as the *dropout mask* [136]). The dropout mask is created using the random variables drawn from the Bernoulli distribution, i.e.,

$$\mathbf{m} = Bernoulli(p), \tag{3.19}$$

where $P(m = 1) = 1 - p$, and $P(m = 0) = p$. Hence, the output at the i^{th} layer is given by:

$$\mathbf{h}_{i+1} = \mathbf{m} \odot f(\mathbf{W}_i^T \mathbf{h}_i + \mathbf{b}_i). \tag{3.20}$$

Hence, dropping out a neuron in DNN with probability, p means that a neuron is dropped out, and its output is set to zero irrespective of whatever the input is given. On the other hand, it will keep the neurons with $1 - p$ probability in the network. We have taken dropout probability of 0.3 as recently suggested in the area of speech recognition [137, 138]. Once the neuron is dropped out, it will not be able to contribute in forward and backward pass of the backpropagation algorithm. Every time a neuron is dropped out, it is like training a new DNN and hence, dropout can be thought as the average result for the entire ensemble of DNNs than a single DNN [136]. Figure 3.10 shows schematic representation of the proposed DNN with dropout.

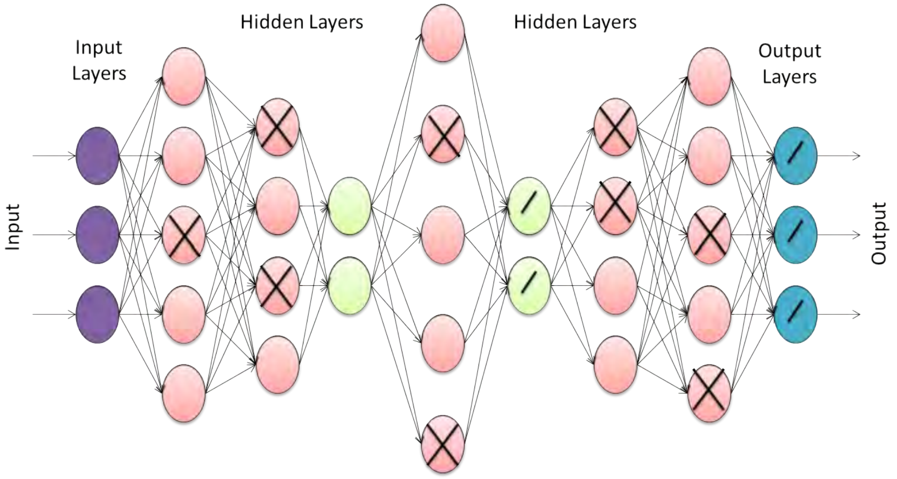


Figure 3.10: Architecture of the proposed network with dropout. Neurons with slanted lines are having linear activation function, and the rest are having non-linear activation function. After [4].

3.5.4 Choice of Activation Functions

Early DNN used sigmoid or tanh nonlinear activation until the Rectifier Linear Unit (ReLU) was proposed [98]. It has been empirically shown that with the ReLU activation function, DNNs can be trained without the need of pre-training in other areas of speech processing apart from the VC [139]. Hence, we propose to use ReLU and its recent variants, such as Leaky ReLU (i.e., LReLU) and Exponential Linear Unit (ELU) for avoiding the pre-training in the area of VC. The activation functions are defined in Table 3.1. Here, u controls the slope of negative part.

Table 3.1: Definition of activation functions. After [1–3]

	ReLU	LReLU	ELU
$f(x)$ for $x \geq 0$	x	x	x
$f(x)$ for $x < 0$	0	$u \cdot x$	$u \cdot (e^x - 1)$

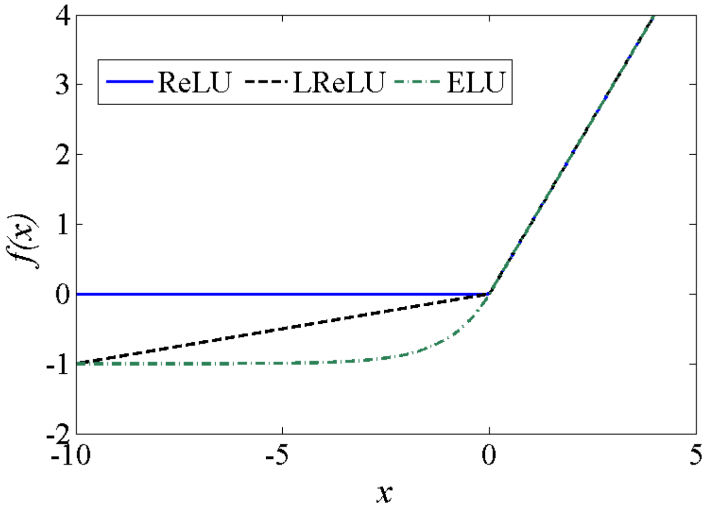


Figure 3.11: Piecewise linear activation functions, where input at neuron $x \in \Re$, and $f(x)$ is the activation function. After [4].

Figure 3.11 shows the three piecewise linear activation functions. Here, LReLU was plotted by taking $u = 0.1$, and for ELU, $u = 1$ is taken as suggested in [2], [3], respectively. The key advantage of the ReLU, LReLU, and ELU is that they do not face gradient vanishing problems that is faced by the sigmoid and tanh [3]. Furthermore, computations of these activations are simpler, which results into speed up in the training and faster convergence. In addition, they generalize the DNN better, i.e., they can predict the values more accurately for the unseen data. Moreover, sigmoid activation is easier to saturate and hence, derivative of the input is almost zero once the sigmoid reaches to the any side of the plateau [98]. However, ReLU saturates only when the input is less than zero. In addition, LReLU reduces this saturation regions due to its non-zero behavior with the negative input. Furthermore, ELU also saturate with the negative value in presence of smaller input, which helps in decreasing forward propagation variations [3]. Recently, effectiveness of ELU (in terms of faster convergence, and better generalization of networks) over ReLU and LReLU was shown in the areas of image processing [3].

3.5.5 Optimization

One of the key challenges of the SGD was that it requires a proper selection of learning rate. For example, too small value of learning rate will lead to the slower convergence, and higher learning rate can lead to miss the true convergence. In addition, the SGD applies the same learning rate to all the parameters. The Adam optimizer computes an individual adaptive learning rates for different parameters from the estimates of first, and second moments of the gradients [140]. The name of Adam is derived from the adaptive moment estimation [140]. The Adam optimization has several advantages, such as the magnitudes of the parameter updates are invariant to rescaling of the gradient, and its step sizes are approximately bounded by the step size of hyperparameters. It also does not require a stationary objective, and works well with the sparse gradients, and it naturally performs a form of step size annealing. Due to the use of bias correction along with the first, and second-order moments of the gradient terms, the Adam optimization was shown to perform better than the SGD-based methods [140]. Recently, its convergence characteristics were also discussed [141].

3.5.6 Xavier Initialization

Proper initialization of the random weights play a key role in the training of the DNN [98]. For example, variance of the input start diminishing as it passes through each layer, if the weights are small. Hence, the inputs will not be useful during the training. Similarly, the variance of input data start increasing as it passes through each layer, if the weights are large. Hence, the inputs will explode, and will not be useful either. To tackle the issues of initializing a DNN with an arbitrary random weights, Xavier initialization technique was proposed in [133]. It ensures that the variance of the weights remains the same as it passes through each layer. This is achieved by initializing the weights from the Gaussian distribution with zero-mean and variance of $1/I$, where I is the number of input neurons [133]. In this work, we also compared all the results w.r.t. the random, and Xavier initialization techniques.

3.5.7 Experimental Results

In this work, both VC Challenge (VCC) 2016, and VCC 2018 databases have been used to build VC systems [118,119]. *25*-D Mel Cepstral Coefficients (MCCs) (including the 0^{th} coefficient), and *1*-D F_0 for each frame (having *25* ms frame duration, and *5* ms frame shift) have been extracted. We have built the VC systems for all *25*, and *16* speaker-pairs given in the VCC 2016 and VCC 2018 databases, respectively.

Table 3.2: Descriptions of VC systems. After [4]

System	Pre-training	Opt.	Activation	Dropout
A (Baseline [47])	✓	SGD	Sigmoid	×
B	×	SGD	Sigmoid	×
C	×	SGD	ReLU	×
D	×	SGD	LReLU	×
E	×	SGD	ELU	×
F	×	SGD	Sigmoid	✓
G	×	Adam	Sigmoid	×
H	×	Adam	Sigmoid	✓
I	×	Adam	ReLU	✓
J	×	Adam	LReLU	✓
K	×	Adam	ELU	✓
L	✓	Adam	Sigmoid	✓
M	✓	Adam	ELU	✓
N	✓	SGD	ELU	✓

Opt.: Optimization, Here ✓ indicates technique (pre-training, dropout) is used; × indicates it is not used

The Dynamic Time Warping (DTW) algorithm was applied for the alignment task. The number of training utterances have been varied from, *n=10, 20, 40, 100,* and *150*. Four speakers' data from the CMU-ARCTIC database has been taken for the pre-training. We employ exactly the same architecture given in [47] for our baseline DNN system. We employ different optimization algorithm, nonlinear activation function, and the dropout techniques in number of combinations w.r.t. the baseline VC systems. In this work, we used $u = 0.01$ for LReLU, and $u = 1$ for ELU nonlinear activation function. We used Adam optimization with the $\beta_1 =$

0.9, and $\beta_2 = 0.999$. The learning rate, and the number of epochs were chosen from [47]. Mean-variance (MV) transformation is used for the F_0 (i.e., fundamental frequency) transformation. The AHOCODER is used for the analysis-synthesis [67]. The description of the developed VC systems is given in Table 3.2.

3.5.7.1 Objective Evaluation

For objective evaluation, we have selected the state-of-the-art Mel Cepstral Distortion (MCD) measure [16]. Figure 3.12 shows the average MCD for all the systems developed using 25 speaker-pairs along with 95 % confidence interval. It can be seen that the system A (i.e., baseline with pre-training) is having lower MCD value than the system B (i.e., baseline without pre-training). This clearly indicates that for a given baseline architecture [47], the pre-training is indeed helping to achieve lower MCD value. The MCD for system B is further reduced with the use of advanced activation functions (as shown for systems C, D, and E). Furthermore, systems G to K (i.e., systems with Adam and dropout and no pre-training) are able to perform equal or better compared to the baseline. This also shows the significance of Adam optimization over SGD. Moreover, it can also be observed that with SGD, and Adam system with ELU is performing better compared to all other activations.

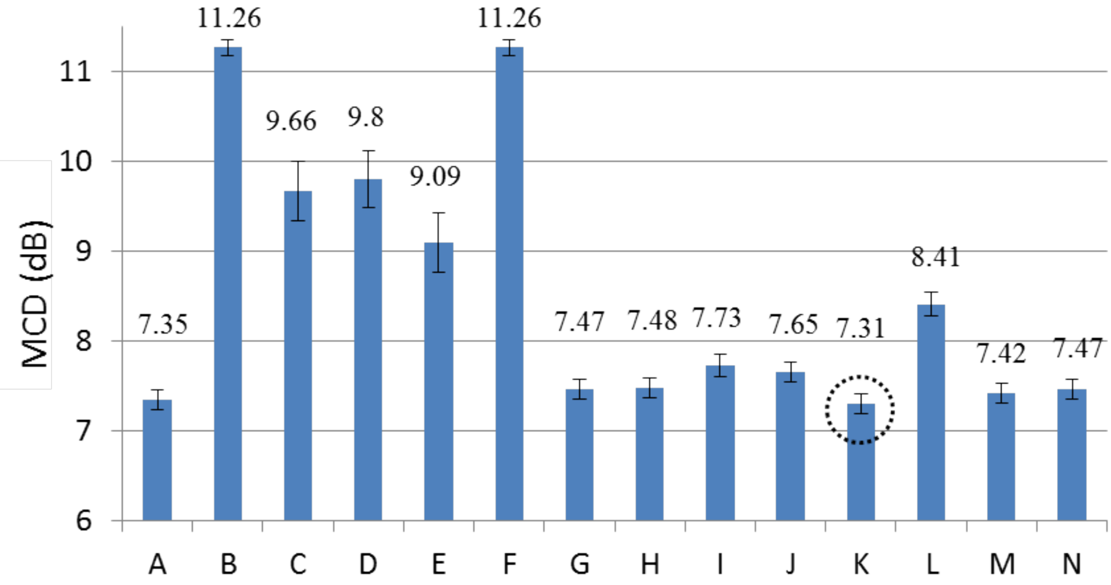


Figure 3.12: The MCD analysis for various VC systems developed on VCC 2016 database. Dotted circle indicates relatively better performing proposed VC system. After [4].

To further investigate the effectiveness of pre-training, we also develop the systems L to N. In the case of pre-trained network, further reduction in the MCD for the system with Adam over SGD can be clearly seen in Figure 3.12 w.r.t. the activation function ELU, and the dropout. Furthermore, reduction in the MCD can be clearly seen for the system M w.r.t. the system L, which is solely due to the ELU activation function. Overall, the system K is performing better w.r.t. the baseline, and other pre-trained network. Hence, the proposed network can be used to overcome the need of pre-training in the VC.

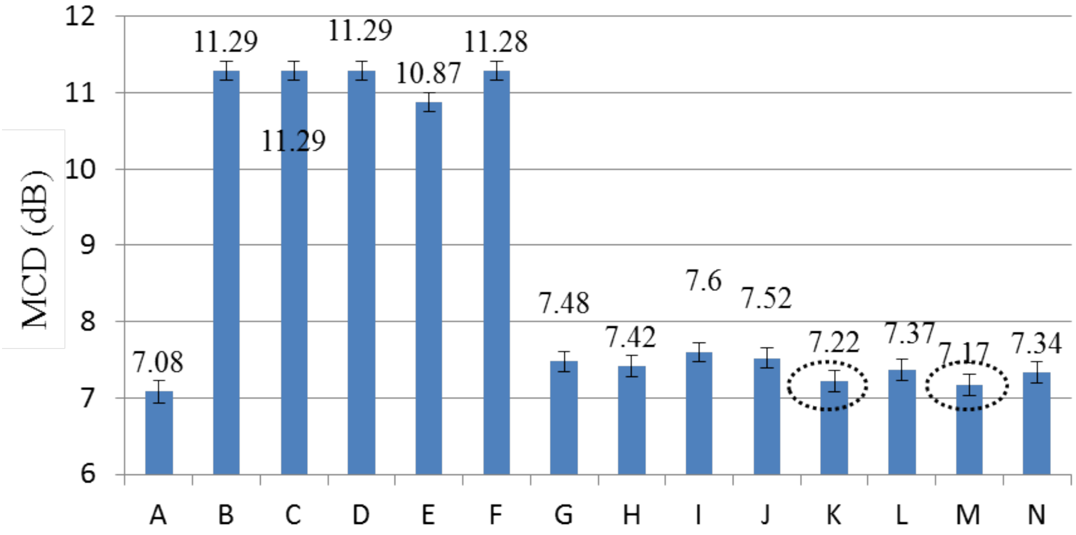


Figure 3.13: The MCD analysis for various VC systems developed on VCC 2018 database. Dotted circle indicates relatively better performing proposed VC system. After [4].

Figure 3.13 shows the average MCD for all the systems developed on the Hub task of VCC 2018 using 16 speaker-pairs along with 95 % confidence interval. Here, average MCD is calculated for all the VC systems developed on the 25 and 16 speaker-pairs in VCC 2016 and VCC 2018 databases, respectively. It can be seen that the system K without pre-training is performing relatively the *best* among all the systems. Furthermore, it can be seen that the system K is consistently performing better across the number of training utterances. In particular, system K with only ten utterances in training (where the possibilities of overfitting is higher) is also performing better than the baseline system A. This may be due to the fact that the dropout prevents overfitting in the DNN by means of stochastic regularization.

Figure 3.14 and Figure 3.15 shows the comparison of the Xavier initialization w.r.t. the random initialization on the VCC 2016 and VCC 2018 databases, respectively. The effectiveness of the Xavier initialization can be seen on both the databases. Hence, the Xavier initialization can also be useful to overcome the need of pre-training in addition to the other proposed modifications. We can clearly see that there is a significant improvement in the performance of System C, D, and E with the Xavier initialization on both the databases. It is possibly due to the fact that the fixed variance in the Xavier initialization at each layers may help the inputs to not getting explode and hence, resulting in the better performance [133].

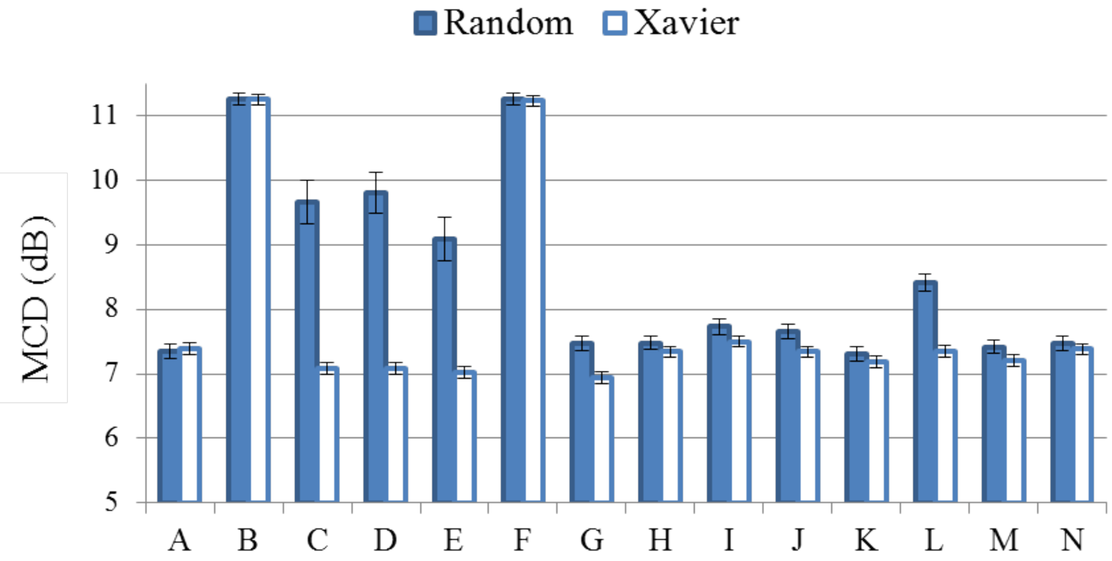


Figure 3.14: Comparison of random initialization *vs.* Xavier initialization on the VCC 2016 database. After [4].

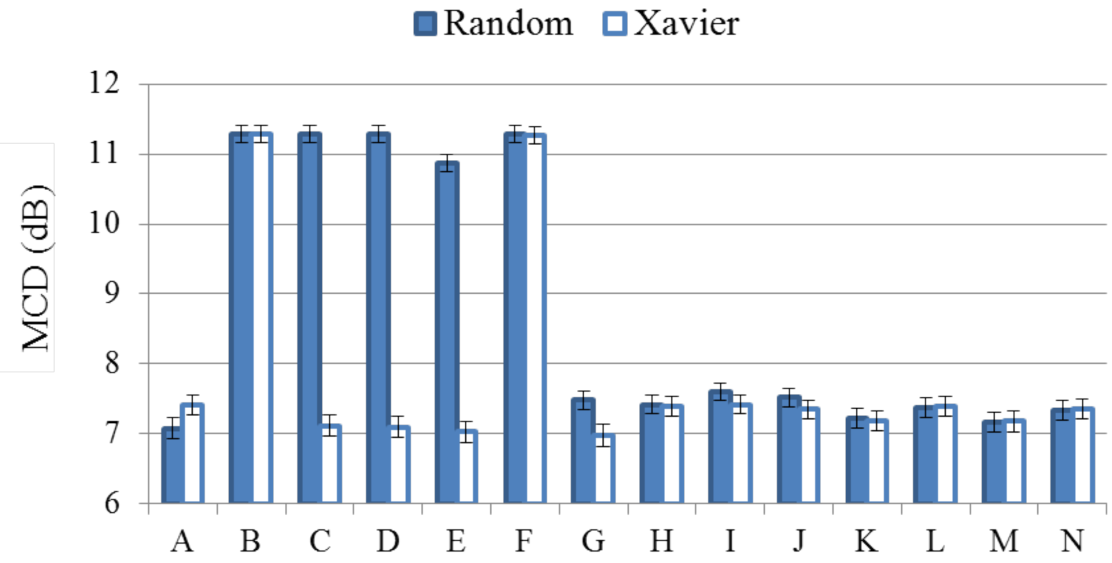


Figure 3.15: Comparison of random initialization *vs.* Xavier initialization on the VCC 2018 database. After [4].

3.5.7.2 Subjective Evaluation

To measure both the speech quality, and the Speaker Similarity (SS) of converted voices, Mean Opinion Score (MOS) tests have been taken. The subjective tests were taken from the 14 subjects (2 female and 12 male with no known hearing impairments, and with the age variations between 21 to 30 years) from total 252 samples. In the MOS test, subjects were asked to evaluate randomly played utterances for the speech quality and SS. For speech quality, subjects were asked to rate the converted voice on the scale of 1 (i.e., very bad) to 5 (i.e., very good). Similarly, for the SS, subjects were asked to rate the converted voice in terms of SS on the scale of 1 (not at all the target speaker) to 5 (exactly the same as that of the target speaker).

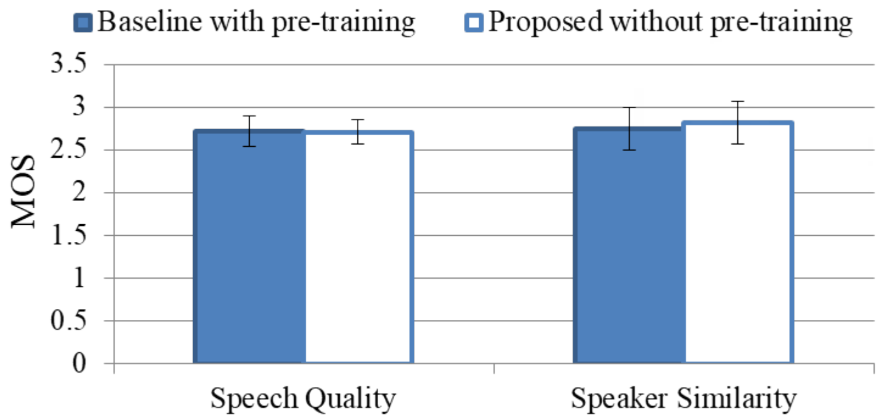


Figure 3.16: The MOS analysis along with *95* % confidence interval on VCC 2016 database. After [4].

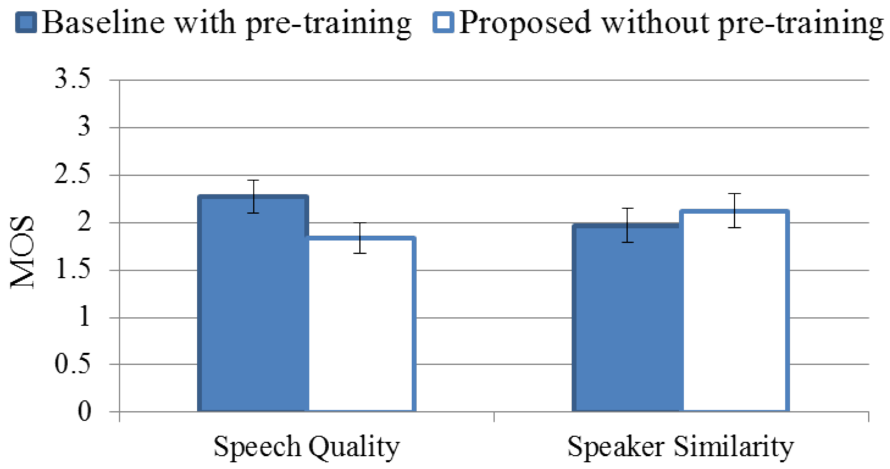


Figure 3.17: The MOS analysis along with *95* % confidence interval on VCC 2018 database. After [4].

The result of the MOS test, for the VCC 2016 and VCC 2018, is shown in the Figure 3.16, and Figure 3.17 along with 95 % confidence intervals, respectively. It is clearly visible from the Figure 3.16 and Figure 3.17 that the proposed system K without pre-training is performing comparable, and slightly better w.r.t. the baseline system with pre-training in the MOS tests for speech quality and speaker similarity, respectively. This indicates that the need of pre-training for the DNN can be reasonably avoided by using our proposed system.

3.6 Proposed MMSE regularized Generative Adversarial Network (GAN)

The traditional deep learning architectures (such as Deep Neural Network (DNN)) are trained on the Maximum Likelihood (ML)-based optimization techniques. GANs provide an alternative to the ML-based optimization criteria [41]. The ability of GAN in modeling deep representation, and learning a suitable mapping function has shown a significant performance improvement in various speech applications, such as Speech Enhancement (SE) [48, 142, 143], Voice Conversion (VC) [11, 103, 107, 144], and Speech Synthesis (SS) [145]. The DNN-based generative models face difficulties in probability computations as appearing in Maximum Likelihood Estimation (MLE) and the related strategies, which are alleviated by the GAN-based architectures [41]. In particular, GANs model the deep representation by learning the statistical distribution of the training data, and synthesizes the novel samples that closely follows the data distribution [41].

The GANs learn the mapping between the features X from some prior distribution $\mathcal{X}$ to samples Y belonging to the data distribution $\mathcal{Y}$. The generator (G) learns the mapping relationships in an *adversarial* framework along with a discriminator (D). The D network is a binary classifier with input as real samples coming from $\mathcal{Y}$, and the estimated samples generated by the G network. The adversarial characteristics of the GAN force the D network to maximize the likelihood of the samples coming from $\mathcal{Y}$ as real, whereas minimizing the likelihood of the generated samples coming from the model distribution $\hat{\mathcal{Y}}$ as fake. The adversarial characteristics force the G network to generate the realistic samples that closely follow $\mathcal{Y}$, essentially developing the *Nash equilibrium*, and leaves the D network unable to differentiate between $\mathcal{Y}$, and $\hat{\mathcal{Y}}$ [41]. At the Nash equilibrium, further improvement in discriminator will not help to the generator and vice versa [41]. Schematic representation of GAN architecture is shown in Figure 3.18. This objective function can be formulated as [41]:

$$\begin{aligned} \min_{D} V(D) &= -\mathbb{E}_{Y\sim\mathcal{Y}}[\log D(Y)] - \mathbb{E}_{X\sim\mathcal{X}}[\log(1 - D(G(X)))], \\ \min_{G} V(G) &= -\mathbb{E}_{X\sim\mathcal{X}}[\log D(G(X))], \end{aligned} \tag{3.21}$$

where $\mathbb{E}_{X\sim\mathcal{X}}$ denotes the expectation over all the samples X coming from the distribution $\mathcal{X}$. The task of discriminator is to give zero probability for the generated output (i.e., $D(G(X))$) by minimizing the $-\mathbb{E}_{X\sim\mathcal{X}}[\log(1 - D(G(X)))]$. Similarly, D's decisions over real data should be accurate by minimizing $-\mathbb{E}_{Y\sim\mathcal{Y}}[\log D(Y)]$. On the other hand, the generator is trained to increase the chances of D producing a high probability for a converted features, thus to minimize $-\mathbb{E}_{X\sim\mathcal{X}}[\log D(G(X))]$.

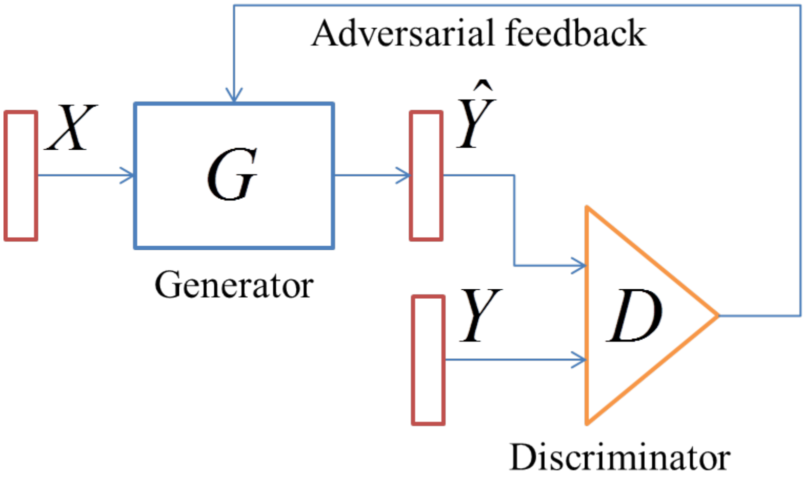


Figure 3.18: Schematic representation of vanila GAN architecture. After [23]. Here $\hat{Y} = G(X)$.

Research Issue: The key limitation of the vanilla GAN-based system is in generating the samples that may not correspond to the given input.

Proposed Solution: The possible solution for this is to force the G network to learn the representation corresponding to the given input. To that effect, we recently proposed to use Minimum Mean Squared Error (MMSE) as a *regularizer* to the vanilla GAN (and will call it as MMSE-GAN), for the speech enhancement, and the NAM2WHSP tasks [23, 48]. The MMSE regularization to the G network objective function can map the estimated, and the ground truth representation that adversarially minimizes the distributional divergence between the model, and the data distribution. The modified objective function can be written as [23]:

$$\min_{G} V(G) = -\mathbb{E}_{X\sim\mathcal{X}}[\log(D(G(X)))] + \frac{1}{2}\mathbb{E}_{X\sim\mathcal{X},Y\sim\mathcal{Y}}[\log(Y) - \log(G(X))]^2. \tag{3.22}$$

3.7 Proposed MMSE regularized Discover Generative Adversarial Network (DiscoGAN)

Recently proposed DiscoGAN can easily learn the cross-domain relationships that are hidden structure in given data [146]. When the mapping to be found between entities are coming from different domains, then it can be represented as cross-domain mapping. This mapping technique is proposed to find the mapping between whispered and normal speeches and the NAM speech and the whispered speech, respectively. Though, NAM, whisper and normal speech are mode of communications, they are different from speech production and perception viewpoints [24]. Hence, finding mapping between NAM, whispered and normal speech can be considered as cross-domain mapping task. More details of the differences between these entities are presented in Chapter 8.

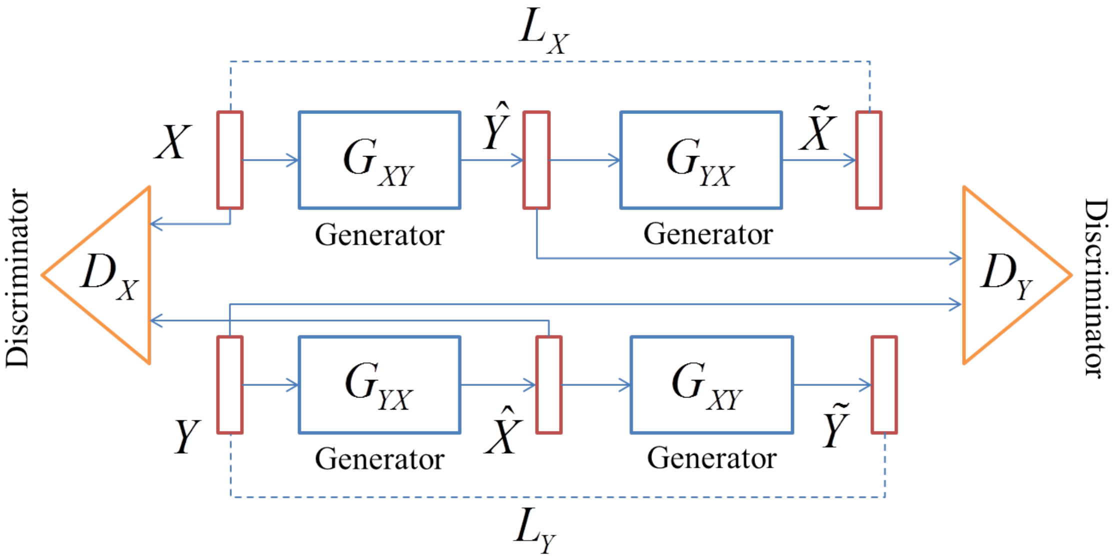


Figure 3.19: Schematic representation of the proposed DiscoGAN architecture. After [24].

In particular, we extended the MMSE-GAN in the framework of Discover GAN (i.e., MMSE DiscoGAN) to learn the cross-domain relations. In particular, we extended the MMSE-GAN via MMSE DiscoGAN by including two generators, G_{XY} and G_{YX} (as shown in Figure 3.19). The G_{XY} mainly converts the X into the $\hat{Y}$, such that $\hat{Y}$ is indistinguishable from the Y. Our model also contains two discriminators, D_X and D_Y. The discriminator D_X attempts to distinguish between X from a distribution $\mathcal{X}$ and the $\hat{X} = G_{YX}(Y)$ obtained by converting the Y from a distribution $\mathcal{Y}$ via the generator G_{YX}. Similarly, D_Y performs an analogous operation for the Y. The minimization of the reconstruction loss (on conversion of

X into Y) enforces the generated $\tilde{X}$ to be as close to X, and the converted $\hat{Y}$ to be as close to Y, by optimization of the MMSE-GAN loss function. These two properties are explored to encourage the one-to-one mapping between the cross-domains [24].

Since our task is to map the X to the Y, we rely on the regularized adversarial objective function, which can be mathematically formulated as [24, 48],:

$$L_{G_X} = -E_{X\sim\mathcal{X}}[\log(D_Y(G_{XY}(X)))] + \frac{1}{2}E_{X\sim\mathcal{X},Y\sim\mathcal{Y}}[\log(Y) - \log(G_{XY}(X))]^2,$$

$$L_{D_Y} = -E_{Y\sim\mathcal{Y}}[\log(D_Y(Y))] - E_{X\sim\mathcal{X}}[\log(1 - D_Y(G_{XY}(X))]. \tag{3.23}$$

The MMSE-GAN is optimized by minimizing the two adversarial loss functions, namely, L_G (generator loss), and L_D (discriminator loss) as shown via eq. (3.23). The training process for the MMSE DiscoGAN is almost similar to the MMSE-GAN [24, 48]. Here, G_{XY}, G_{YX}, D_X, and D_Y must be jointly trained, with one significant modification of including two reconstruction losses, L_X and L_Y. This can be mathematically represented as:

$$L_X = d((G_{YX}(G_{XY}(X)), X), \text{ and } L_Y = d((G_{XY}(G_{YX}(Y)), Y). \tag{3.24}$$

These reconstruction losses (given in eq. (3.24)) satisfy our requirement that G_{XY} and G_{YX} must be inverse of each other to the extent possible, i.e., for any X, $\tilde{X} = G_{YX}(G_{XY}(X))$ must be close to the X, and similarly, for any Y. Ideally, the equality of $\tilde{X}$ and X (i.e., $\tilde{X} = X$) should hold. However, this is difficult to optimize. For this reason, the distances L_X, and L_Y are minimized using the *Huber loss* as the metric function [146]. Hence, the generator loss for G_{XY} can be defined as [Th 12]:

$$L_{G_{XY}} = L_X + L_{G_Y}, \tag{3.25}$$

where L_{G_Y} can be defined using eq. (3.23). The generator loss G_{YX} can also be defined in the similar way. Hence, the total generator loss is $L_{G_{XY}} + L_{G_{YX}}$, and the total discriminator loss is $L_{D_X} + L_{D_Y}$, where L_{D_X} and L_{D_Y} can be defined using eq. (3.23).

Experimental results for the MMSE-GAN and MMSE-DiscoGAN-based architectures for the VC task and the cross-domain conversion tasks are discussed in the Chapter 6 and Chapter 8, respectively.

3.8 Chapter Summary

In this chapter, we briefly summerized the contributions of the thesis from mapping perspective in the context of VC task. In particular, we first discuss the details of GMM-based VC along with its strengths and limitations. Next, we presented a novel AS technique for BLFW-based VC, which exploits the advantage of GMM-based VC for better speech quality compared to the state-of-the-art AS technique. Later, we also presented Neural Network-based mapping techniques. In addition, we also presented empirical analysis to show that with the proposed DNN architecture, need of pretraining can be avoided in the context of VC without compromising the speech quality and speaker similarity of the converted voice. Finally, the proposed MMSE regularized GAN and DiscoGAN architectures have been discussed.

Even though alignment step should come first than the mapping function for VC, we assume here that we had access to the time-aligned corresponding feature-pairs from the source, and the target speakers. Once, the aligned corresponding pairs are available, the task of mapping function is the same as supervised learning problem. However, obtaining such corresponding pairs is a different task for parallel as well as non-parallel VC. Now, in the next two chapters (Chapter 4 and Chapter 5), we will focus on how to obtain such corresponding pairs in the case of parallel and non-parallel VC tasks.

Chapter 4

Alignment Strategies for Parallel VC

4.1 Introduction

In the last chapter, we presented several mapping techniques that can be applied once the corresponding feature pairs are obtained after the alignment step. In this chapter, we present the alignment techniques to obtain corresponding feature pairs in the case of parallel VC. In parallel VC task, the spectral features from both source and the target speakers must be time-aligned during training due to speaking rate variations across the speakers (i.e., interspeaker variations), and speech rate variations within the speaker (i.e., intraspeaker variations). Wrongly-aligned pairs will affect the learning of the *mapping* function, which in turn will deteriorate the quality of the converted voices [39]. Hence, the alignment is one of the key steps in the parallel data VC task. In the case of a parallel data VC, a Dynamic Time Warping (DTW) algorithm is used for the alignment. Section 4.2 describes the DTW algorithm, and its limitations. Section 4.3 presents novel outliers removal strategies as a pre-processing step to overcome the issues related to the alignment in the parallel VC. Section 4.4 summarizes the experimental results in order to show the effectiveness of proposed outlier removal strategies. Finally, we also present the comparison between the proposed outlier removal approach, and the robust alignment strategies in Section 4.5

4.2 Dynamic Time Warping (DTW) algorithm

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, and $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_M\}$ be the short-time spectral feature vector sequences corresponding to the training utterances, where N and M are the lengths of the two feature vector sequences, respectively. Hence, inputs to the DTW are the $\mathbf{X}$ and $\mathbf{Y}$ spectral feature sequences of different lengths from source and the target speakers, respectively. Output will be the warping path (i.e., the order in which we want to place features from $\mathbf{X}$ and $\mathbf{Y}$ such that both will be having the same length). $\mathbf{X}_{w_x(k)}$ and $\mathbf{Y}_{w_y(k)}$ are the time-aligned feature sequences obtained after the DTW for the feature sequences $\mathbf{X}$ and $\mathbf{Y}$, respectively. In particular, $w_x(k)$ and $w_y(k)$ contains the indices of frames in a specific order obtained after the DTW in order to place features frame-by-frame to get aligned spectral feature sequences of the same length for $\mathbf{X}$ and $\mathbf{Y}$ (i.e., $\mathbf{X}_{w_x}$ and $\mathbf{Y}_{w_y}$), respectively. An illustration of the nonlinear warping path obtained after DTW for an utterance spoken by both the source and the target speakers is shown in Figure 4.1.

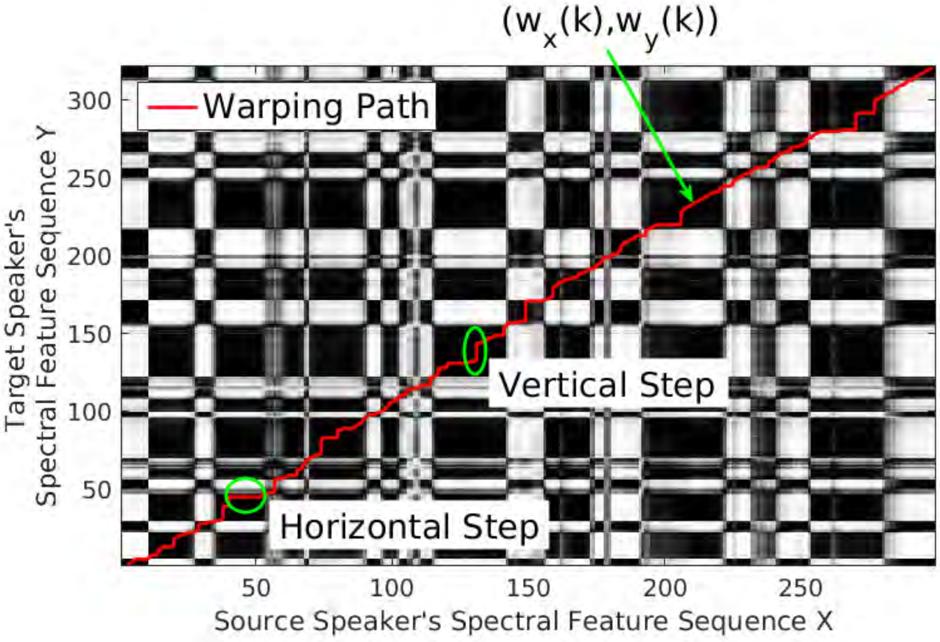


Figure 4.1: Illustration of the nonlinear warping path obtained after DTW. After [5].

The main objective of the DTW is to find an optimal nonlinear warping path between the two spectral sequences, $\mathbf{X}$ and $\mathbf{Y}$. The DTW minimizes the overall distance $d(\mathbf{X}, \mathbf{Y})$, which is the sum of a local distance $d(\mathbf{x}_{w_\mathbf{x}(k)}, \mathbf{y}_{w_\mathbf{y}(k)})$ computed over the warping path $w_\mathbf{x}$, and $w_\mathbf{y}$. Hence, the DTW can be expressed as [73]:

$$DTW(\mathbf{X}, \mathbf{Y}) = \min_{w_\mathbf{x}, w_\mathbf{y}} \sum_{k=1}^{K} d(\mathbf{x}_{w_\mathbf{x}(k)}, \mathbf{y}_{w_\mathbf{y}(k)}), \tag{4.1}$$

where $d(.)$ is the Euclidean distance, and K is the total length of the warping path. The objective function given by eq. (4.1) is optimized using the following three standard constraints (as given in [73]):

1. **Boundary constraint:** $(\mathbf{x}_{w_\mathbf{x}(1)}, \mathbf{y}_{w_\mathbf{y}(1)}) = (\mathbf{x}_1, \mathbf{y}_1)$, and $(\mathbf{x}_{w_\mathbf{x}(L)}, \mathbf{y}_{w_\mathbf{y}(L)}) = (\mathbf{x}_N, \mathbf{y}_M)$, i.e., the warping path starts and stops at the first and the last spectral feature vectors from the source and the target speakers', respectively.

2. **Monotonicity constraint:** $w_{\mathbf{x}}(1) \leq w_{\mathbf{x}}(2) \leq \ldots w_{\mathbf{x}}(L)$, and $w_{\mathbf{y}}(1) \leq w_{\mathbf{y}}(2) \leq \ldots w_{\mathbf{y}}(L)$, i.e., the warping path must not have a negative slope.

3. **Step size constraint:** $(w_{\mathbf{x}}(k+1), w_{\mathbf{y}}(k+1)) - (w_{\mathbf{x}}(k), w_{\mathbf{y}}(k)) = (1,0), (0,1), (1,1)$, i.e., at any point in the warping path, transitions to the current node are allowed from the set of predefined predecessors node such that the step size constraint is satisfied.

4.2.1 Limitations of DTW and Sources of Outliers

Due to the boundary constraint, in the presence of frames corresponding to the silence regions (either at the beginning or at the end of the source and/or the target speaker's speech signal), warping path may generate speech-silence pairs (i.e., speech *vs.* non-speech pairs). In addition, due to the monotonicity constraint, speech-silence pairs may also be created due to the presence of silence regions between various words or phrases in an utterance. Furthermore, due to the step size constraint, the warping path will no longer be linear and hence, the nonlinear warping path of the DTW will produce one-to-many and many-to-one aligned pairs due to vertical and horizontal steps (as shown in Figure 4.1). Here, black and white colors in the background of Figure 4.1 indicates the value of a distance matrix, which is calculated frame-by-frame among the frames of **X** and **Y**. In particular, black color indicates the lowest distance, and the white color represents the highest distance.

Many of these unwanted pairs may lead to the data points that do not follow the overall trend of the data and hence, such issues in the DTW alignment can be considered as one of the major sources of *outliers* in the context of VC. Furthermore, we are trying to minimize the distance between spectral features for a given speaker-pair with above mentioned constraints. Though human speech production mechanism is the same across humans, there is a significant difference in the spectral representations across the speakers. This is primarily due to the differences in vocal tract system (shape and size) and excitation source (differences in size of glottis, mass of vocal folds, tension in vocal folds and hence, the manner in which glottis opens or closes, i.e., the glottal activity) across the speakers [17]. These differences are more in the case of intergender speaker-pairs than the intragender speaker-pairs. Hence, when we try to minimize the distance between the spectral features without considering these differences for the alignment, some corresponding pairs are obtained that are inconsistent with the rest of the data and are considered as *outliers*. In this thesis, we propose to remove such outliers before learning the mapping function.

The removal of outliers will help in estimating the correct parameters of the regression during the training. Outliers removal is a very popular in wide varieties of applications, where the outliers occur due to the mechanical faults, changes in a system behavior, human error, instrument errors, etc. [147–151]. In addition, sources of outliers in the speech data due to environmental noise have also been identified [152, 153]. However, in the context of VC task, one of the main causes of the generation of outliers is the issues related to the alignment techniques [39], [49].

4.3 Proposed Outliers Removal Approach for VC

4.3.1 Impact of Outliers on Regression

Outliers tend to shift the mean and scatter (i.e., covariance matrix) of the data away from their true values, which significantly affect the performance of the statistical-based approaches [154]. In this context, we present here a toy example of linear regression to visualize the effect of the presence of outliers on its performance. Here, we have taken the 2-*D* data points (with the outliers), i.e., feature vectors (x_1, x_2) on which a simple linear regression model has been applied. It can be seen from Figure 4.2 that the estimated slope of the regression line has changed significantly due to the presence of the outliers.

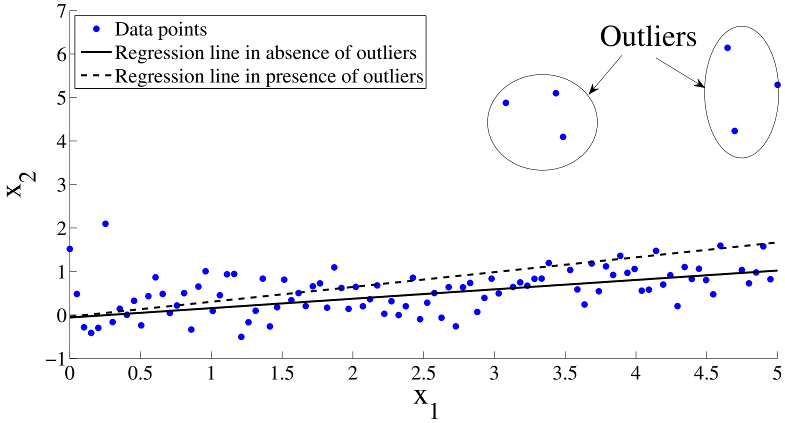


Figure 4.2: Toy example to understand the effect of outliers on the performance of linear regression task. After [5].

In a second toy example, we have taken the 2-*D* data points from the normal distribution to visualize the effect of outliers on the training of GMM. We can see from Figure 4.3 that in the presence of outliers, the estimated mean of the components gets shifted away from its true value. In addition, the estimated covariance of the components is highly affected. Since the GMM-based VC methods are very much dependent on the estimated mean and covariance of the mixture components, the errors in their estimation will affect the performance of the VC system. One of the possible solutions for the above mentioned problem is to increase the number of mixture components in the GMM. However, when dealing with limited training data, the estimation of weights, mean, and covariance matrices for more components will lead to *overfitting*.

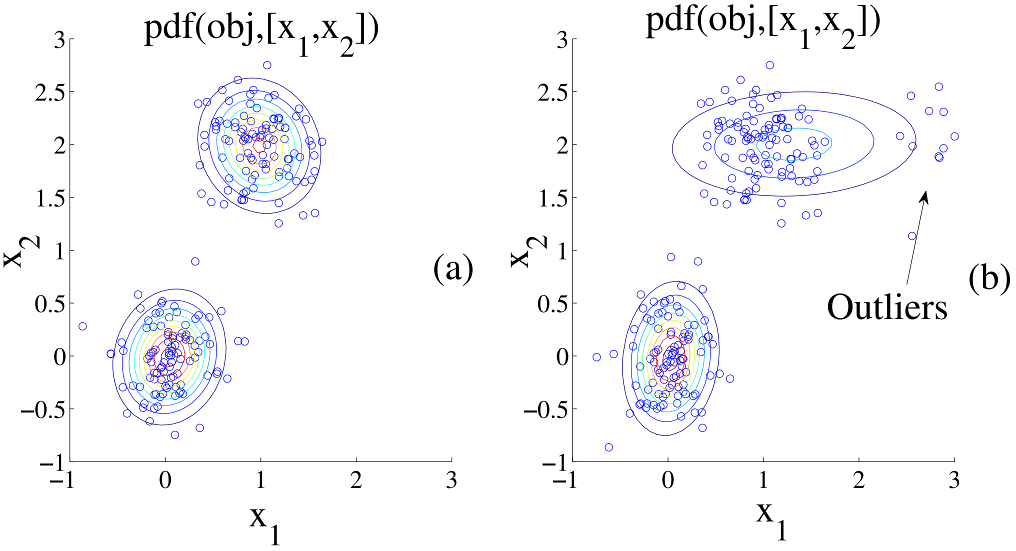


Figure 4.3: Toy example to understand the effect of outliers on the performance of GMM training (a) without outliers, and (b) in the presence of outliers. After [5].

Similarly, in order to understand the impact of outliers in the case of a nonlinear regression task, we have taken 2-*D* data along with the outliers. We have developed a two-layer neural network with Rectifier Linear Unit (ReLU) as an activation function. The DNN is trained using the Adam optimization technique for the two cases, namely, (a) without any outliers, and (b) in the presence of outliers. The trained weight matrices of both the cases are shown in Figure 4.4. We can clearly see significant differences between the learned weights by comparing the two cases. This difference has also affected the performance of the nonlinear regression (as shown in Figure 4.5). We can clearly see that the shape of the nonlinear regression curve gets shifted upward in the presence of outliers, which affects the prediction in the nonlinear regression task. Due to the adverse effects of outliers on the regression techniques (as shown in this Section), the outlier removal should be considered as an essential pre-processing step in various state-of-the-art VC techniques.

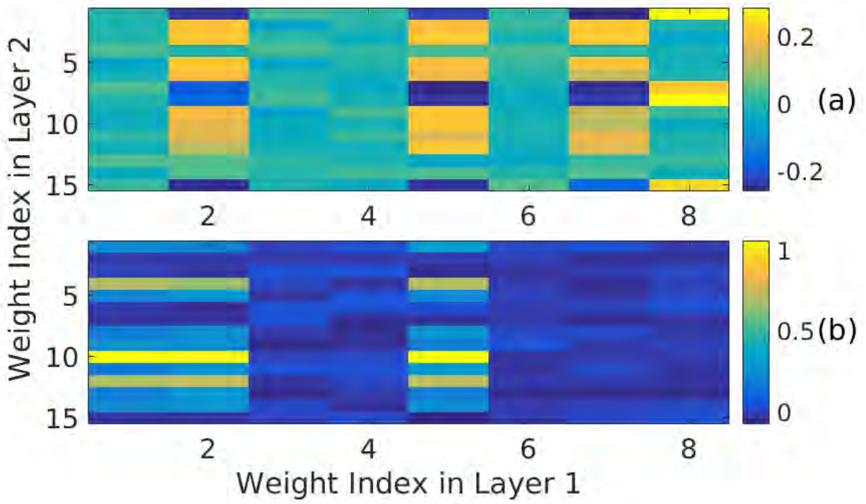


Figure 4.4: Visualization of weight matrix of DNN that contains two hidden layers for the toy data (a) without outliers, and (b) in the presence of outliers. After [5].

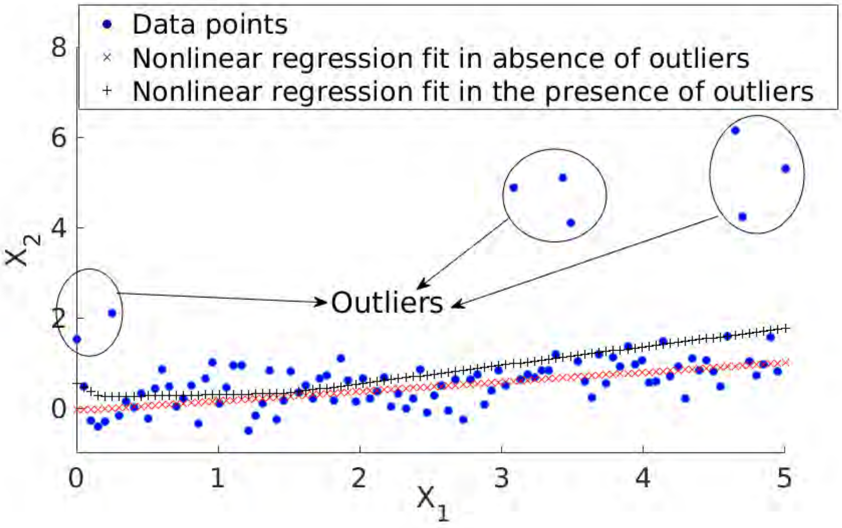


Figure 4.5: Toy example to understand the effect of outliers on the performance of nonlinear regression task. After [5].

4.3.2 Proposed Outliers Detection Method

Detecting outliers in the case of *2*-D and *3*-D is relatively easier since one can visualize the data, and detect the data points that are far away from the general trend of the data. However, the outlier detection becomes very challenging in the case of multivariate data with more than three dimensions. One of the possible ways to identify the outliers in a multivariate data case is by calculating the distance of each point to the center of the data. Then, the outliers are the points that are having a larger distance to the center of the data. The distance from a data point $\mathbf{X}$ to a location $\mathbf{Y}$ in the multivariate case can be computed by the square of the Mahalanobis distance (i.e., $d^2_\Sigma(\mathbf{X}, \mathbf{Y})$), which is given by [155]:

$$d^2_\Sigma(\mathbf{X}, \mathbf{Y}) = (\mathbf{X} - \mathbf{Y})^T \Sigma^{-1} (\mathbf{X} - \mathbf{Y}), \tag{4.2}$$

where Σ is an estimated sample covariance matrix of the data. Hence, the outliers are the points $\mathbf{x}_i$ that are having $d_\Sigma(\mathbf{x}_i, \bar{\mathbf{X}})$ greater than the predetermined cut-off. Here, $\bar{\mathbf{X}}$ (i.e., the sample mean of the data), and Σ (i.e., the sample covariance of the data) are strongly *susceptible* to the outliers. Hence, the decision taken based on these estimates will not be able to detect the actual outliers. This is called a *masking* effect in the context of outlier detection [156]. Though these outliers will be at a larger distance from the general trend of the data in a certain projection, identification of such a projection is itself a very challenging task. The PCA can be used to identify the projection by extracting the relevant information in a high-dimensional multivariate dataset.

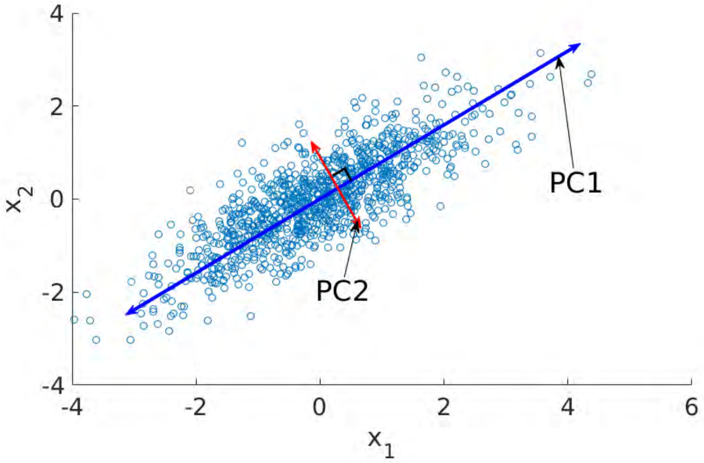


Figure 4.6: Visualization of principal components for 2-*D* data, i.e., $x = [x_1, x_2]$. PC1: First principal component, PC2: Second principal component. After [5].

The dimensionality reduction is achieved by looking at the covariance between the variables and projecting the information on a fewer dimensions that are having the maximum covariances between them. As a result, the Principal Component Analysis (PCA) is often used as a first stage in the statistical analysis of the multivariate data [157]. The standard PCA method first computes an eigenvector having the largest eigenvalue of the covariance matrix of the data. This is done since the dominant eigenvector shows the direction in which the data has the largest variance. After projecting the data on the first component, we decompose the covariance matrix. This matrix is used to compute again the dominant eigenvector for maximizing the covariance for finding the second component. The second component is *orthogonal* to the first one (as shown in Figure 4.6). This procedure continues till we project the data on the pre-selected number of components. The optimum number of components can be computed using cross-validation techniques or selected manually. As mentioned earlier, the first component is found by maximizing the variance and hence, it is more susceptible to include the outliers, and fails to capture the variance of the regular data (i.e., data without outliers). To avoid such limitations, we consider the data obtained, after removing the outliers, to compute the principal components. This method is called the Robust PCA (ROBPCA) [158, 159].

In this thesis, we have used the ROBPCA for dimensionality reduction. The ROBPCA is performed on the concatenated data, $\mathbf{Z}_{n,d} = [\mathbf{X}; \mathbf{Y}]$. Here, n is the number of frames, and d is the number of variables of the concatenated feature vector, $\mathbf{Z}$. First, we apply Singular Value Decomposition (SVD). The SVD is performed on the mean-centered data matrix. Here, the dominant eigenvector of the covariance matrix is not retained, since this would imply that we were performing the standard PCA. We find h least outlying points using the Minimum covariance Determinant (McovD) estimator, and use their covariance matrix to obtain a p-dimensional feature subspace. The number of h least outlying points is calculated as $h = max\{[\alpha n], [(n + p_{max} + 1)/2]\}$, where the default value of p_{max} is 25 (since we have taken *25*-D Mel Cepstral Coefficients (MCC)), and the parameter $\alpha = 0.75$ (empirically chosen for optimal computational cost [158]). Using this set of h data points, we find the location and scatter (i.e., mean and covariance) of the data using the McovD estimator. The objective of this estimator is to find a certain number of observations that have the lowest determinant of their covariance matrix. The McovD is robust to the outliers. Hence, the McovD is given by [160]:

$$McovD = (\hat{\mu}_J, \hat{\Sigma}_J), \tag{4.3}$$

where

$$\begin{aligned} J =& \{\text{set of } h \text{ points: } |\hat{\Sigma}_J| \leq |\hat{\Sigma}_r|, \forall r \text{ s.t. } |r| = h\}, \\ \hat{\mu}_J =& \frac{1}{h}\sum_{i \in J} \mathbf{x_i}, \\ \hat{\Sigma}_J =& \frac{1}{h}\sum_{i \in J} (\mathbf{x_i} - \hat{\mu}_J)(\mathbf{x_i} - \hat{\mu}_J)^T. \end{aligned} \tag{4.4}$$

Here, x_i is the i^{th} data point from the set of h data points. The ROBPCA uses a computationally fast variant of the McovD estimator, called the Fast Minimum Covariance Determinant (FMcovD) algorithm [157]. The eigenvalues and their corresponding eigenvectors of the final robust scatter matrix are calculated. The first p-dominant eigenvectors are retained, and they form the p-dimensional loading matrix, $\mathbf{P}_{n,p}$. This loading matrix, along with the final robust mean obtained earlier, is used to compute the scores using the formula [158]:

$$\mathbf{T}_{n,p} = (\mathbf{Z}_{n,d} - \mathbf{1}_n \hat{\mu}^T)\mathbf{P}_{n,p}. \tag{4.5}$$

Here, $\mathbf{1}_n$ is the notation used for column vector with all n components that are having value 1, and $\hat{\mu}$ is the robust mean obtained using McovD. The score matrix given in eq. (4.5) is used to calculate the score distances that further help us in determining the outliers. In particular, the score distance can be calculated as [158]:

$$SD_i^{(p)} = \sqrt{(\mathbf{t}_i - \hat{\mu})^T (\hat{\Sigma})^{-1} (\mathbf{t}_i - \hat{\mu})}, \tag{4.6}$$

where $\hat{\mu}$ and $\hat{\Sigma}$ are estimated using McovD and SD_i^p is the score distance of the i^{th} frame, when p is the number of components in the ROBPCA [158]. In addition, $\mathbf{t_i}$ is the i^{th} column vector of the score matrix, $\mathbf{T}_{n,p}$. Once the score is estimated, the outliers are detected and removed based on the cut-off from the training data.

4.3.3 Proposed Threshold Selection Method

Selecting an exact threshold (i.e., cut-off) for detecting the outliers is an issue [156, 161]. In most of the cases, it has been assumed in the literature that the score distances follow the normal distribution since the data are assumed to be normally distributed [158]. Hence, the squared Mahalanobis score distance of normally distributed data follows the χ_p^2 distribution approximately. Thus, a cut-off equal to the $\sqrt{\chi_{p,0.975}^2}$ is taken for detecting the outliers [158], where p is the degrees of freedom in a chi-squared distribution. We have kept the value of p equal to the number of components selected in the ROBPCA. Frames that are having score distance more than this cut-off are termed as the outliers and hence, they are removed before the training of VC system. In our earlier work, we selected the same speaker-pair independent cut-off based on a χ_p^2 distribution assumption [49]. However, it has been observed that the chi-squared approximation is a weak statistical assumption [161].

In particular, we observed that the logarithm of score distances follows more closely the normal distribution than the distribution of score distances itself (as shown in Figure 4.7). Hence, we propose to explore various speaker-pair dependent cut-offs for the outlier detection task in this thesis. In particular, we have selected $\mu + 2\sigma$, and $\mu + 3\sigma$ as the cut-offs (where μ and σ are the mean and the variance of the logarithm of the score distances, respectively).

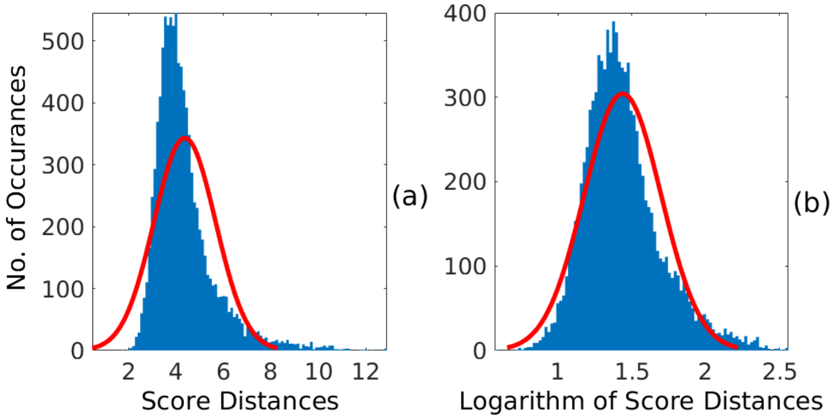


Figure 4.7: Histogram of (a) score distances, and (b) logarithm of score distances with normal probability density function (*pdf*) generated with the same mean and variance. After [5].

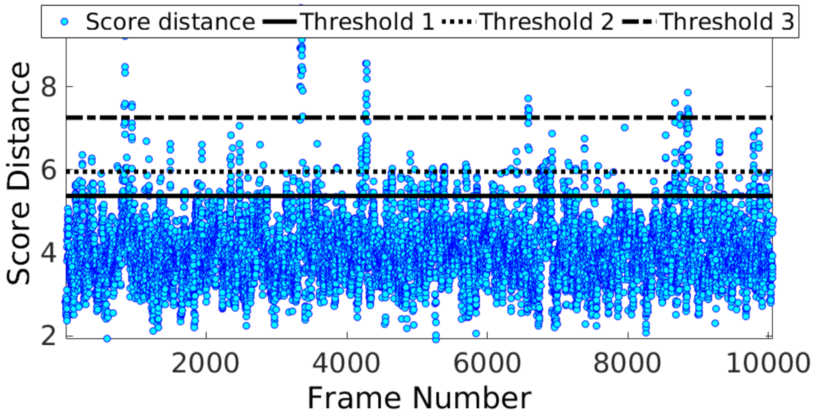


Figure 4.8: Score distance for aligned spectral feature pairs for one of the given source, and target speaker-pair. Threshold 1: $\sqrt{\chi^2_{p,0.975}}$, Threshold 2: $\mu + 2\sigma$, and Threshold 3: $\mu + 3\sigma$. After [5].

The μ, and σ are calculated from the score distances, which are specific to the speaker-pair taken into the consideration. Hence, the proposed cut-offs are speaker-pair dependent. In our earlier work, we selected the same speaker-pair independent cut-off based on χ^2_p distribution assumption [49]. However, it has been observed that the chi-squared approximation is a weak statistical assumption even in the cases where quite a large number of samples are available [161]. In particular, we observed that the logarithm of score distances follow more closely the normal distribution than the distribution of score distances itself as shown in Figure 4.7.

Algorithm 1 Proposed Algorithm for Outlier Removal. After [5]

```
1: Input: Speech utterances from source and the target speakers.
2: Spectral features extraction from both the speakers.
3: Alignment of the spectral features using DTW algorithm.
4: Joint spectral features vector (i.e., $\mathbf{Z} = [\mathbf{X}; \mathbf{Y}]$) formulation.
5: $McovD = (\hat{\mu}_J, \hat{\Sigma}_J)$ estimation for applying ROBPCA on $\mathbf{Z}$
6: Estimation of the first $k$-dominant eigenvectors and the calculation of the score
   matrix, $\mathbf{T}_{n,p} = (\mathbf{Z}_{n,d} - \mathbf{1}_n\hat{\mu}')\mathbf{P}_{n,p}$.
7: $SD_i^{(p)} = \sqrt{(\mathbf{t}_i - \hat{\mu})^T(\hat{\Sigma})^{-1}(\mathbf{t}_i - \hat{\mu})}$, score distance calculations.
8: Select cut-off (such as, $\sqrt{\chi^2_{p,0.975}}$, or speaker-pair specific cut-off) for the outlier
   detection.
9: Remove the frames from the training that are having a score distance more
   than the cut-off.
10: Estimate the mapping or transformation function using various VC techniques
    on remaining frames.
```

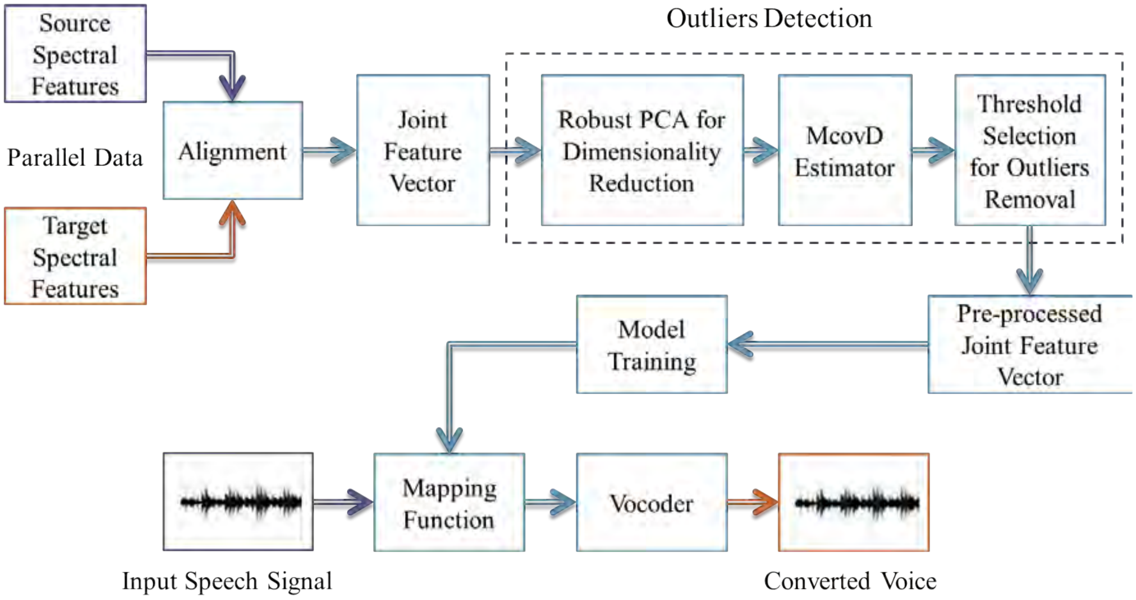


Figure 4.9: A schematic block diagram of the proposed outlier removal approach for parallel data VC. Dotted box indicates a novelty of the work in the outlier detection.

Figure 4.8 shows the scatter plot of the score distances calculated using eq. (4.6) and the various cut-offs for one of the speaker-pairs available for the VC from the CMU-ARCTIC database. The points having a score distance above the cut-off line are removed from the dataset before learning the mapping function during the training for the VC task. In Figure 4.8, Threshold 1: ($\sqrt{\chi^2_{p,0.975}}$) is the speaker-pair independent cut-off, Threshold 2: ($\mu + 2\sigma$), and Threshold 3: ($\mu + 3\sigma$) are the speaker-pair dependent cut-offs. The proposed method for the outlier detection is presented in Algorithm 1. The overall block diagram of the proposed system (as discussed in Section 4.3) is shown in Figure 4.9.

4.4 Experimental Results

4.4.1 Experimental Setup

In this thesis, we have used the CMU-ARCTIC [18] and the VCC 2016 databases [118] for the development of various VC systems in order to investigate effectiveness of outlier removal task. Out of 1132 utterances from each speaker in the CMU-ARCTIC database, we have selected 1078 utterances for training, and 54 for testing. We have built numerous VC systems for *4* and *25* different speaker-pairs from the CMU-ARCTIC and the VCC 2016 databases, respectively. Here, *25*-D Mel Cepstral Coefficients (MCC) (including the *0th* coefficient), and *1*-D fundamental frequency (F_0) were extracted for each 25 ms frame with 5 ms frameshift. The 0^{th} coefficient relates to the absolute energy, and it is mostly ignored in the VC literature. However, one of the baselines that we took for ANN-based VC has, in fact, exploited the 0^{th} coefficient of MCC feature vector [96, 97]. Hence, we adopted the same parameterization and the architecture to compare the effectiveness of the outlier removal. The DTW algorithm has been applied for the alignment task [73]. We have developed VC systems on 4 and 25 speaker-pairs using GMM, ANN, and DNN-based VC techniques on the CMU-ARCTIC and VCC 2016 databases, respectively. The number of training utterances is varied as $\{10, 40, 100, 200, 400, 1078\}$, and $\{10, 20, 40, 100, 150\}$ for CMU-ARCTIC and VCC 2016, respectively. All the VC systems are developed without any pre-processing, and with outlier removal techniques as a pre-processing for the three different proposed cut-offs. Hence, the total number of VC systems for both the database is 1740. In the case of GMM-based VC systems, the number of mixture components is optimized from the set, $m = \{4, 8, 16, 32, 64, 128\}$. The notations used for the developed VC systems are presented in Table 4.1.

Table 4.1: Descriptions of developed VC systems. After [5]

Systems	Descriptions
S1	Without outlier removal
S2	With outlier removal: Speaker-pair independent cut-off $\sqrt{\chi^2_{p,0.975}}$
S3	With outlier removal: Speaker-pair dependent cut-off $(\mu + 3\sigma)$ of logarithm of score distances
S4	With outlier removal: Speaker-pair dependent cut-off $(\mu + 2\sigma)$ of logarithm of score distances

The number of neurons in the ANN-based VC is $25, 50, 50,$ and 25 in the input layer, two hidden layers, and the output layer, respectively [96]. In the case of the DNN-based VC, the number of neurons is $25, 100, 40, 15, 50, 15, 40, 100, 25$ in the input, hidden layers and output layer, respectively. The DAE is used for the pre-training of the DNN. For developing the DNN-based VC systems on the CMU-ARCTIC database, four speaker-pairs from the VCC 2016 database are taken for the pre-training, and vice-versa. The DAE with three autoencoders (AEs) (with size $100, 40,$ and 15) is stacked together, where the first AE is the de-noising AE, which is corrupted with additive Gaussian noise, and later two AEs are the contractive autoencoders [47]. The first AE is trained on the input, and the encoded features obtained from the first AE, are then used for the training of the second AE, and so on. Here, the greedy layerwise training of the DAE is done [135]. The sigmoid function is used as the nonlinear activation function for the hidden layers in both the ANN and DNN-based VC. For the output layer, the linear activation function is used. The ANN and DNN-based VC systems were trained using the SGD algorithm for 1000 epochs with a batch size of 250. The learning rate was set to 0.01 during the training. Out of all the training data, 20% of the data are taken for validation. Mean-variance (MV) transformation is used for F_0 transformation. The detailed objective and subjective analysis of the VC systems are discussed in the following two sub-Sections.

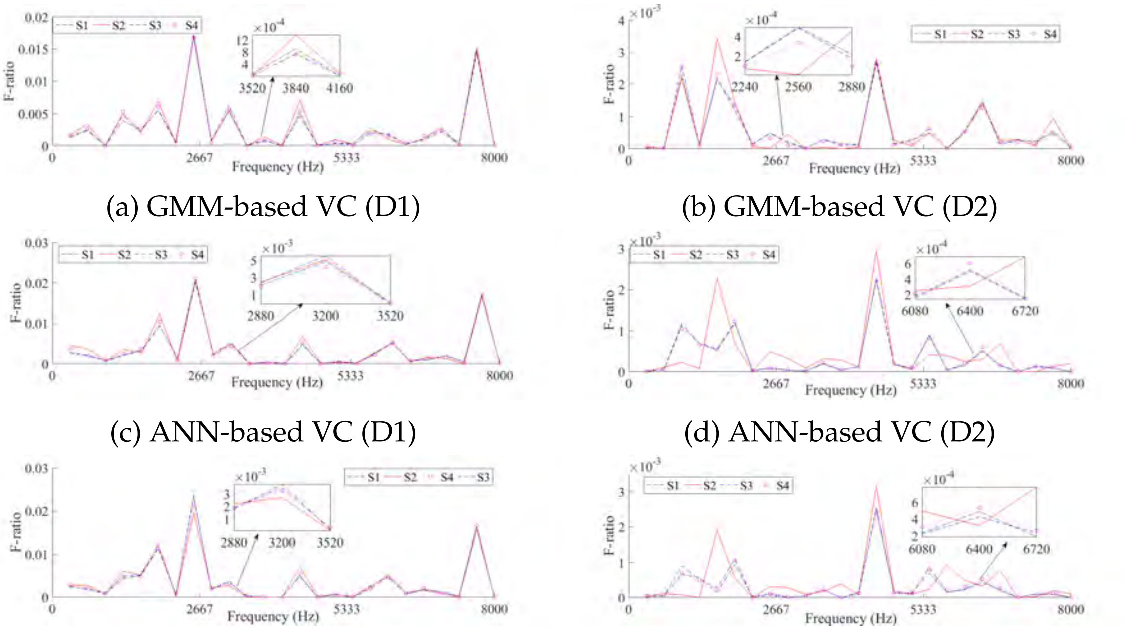
4.4.2 Objective Evaluation

We have used F-ratio-based analysis and state-of-the-art Mel Cepstral Distortion (MCD) for the objective evaluation of the converted voices developed using various VC systems. In pattern recognition, the Fisher's F-ratio is used to measure the discriminative ability of a feature set [162]. In particular, the F-ratio can be used to determine which frequency band is more discriminative to separate the two classes. The F-ratio is a ratio of the inter-class variance to the intra-class variance [162]. Hence, the higher the value of F-ratio means better the separation of the two classes. The F-ratio analysis has been used for the speaker recognition as well [163–165] and recently, in the Spoof Speech Detection (SSD) task as well [166], [167]. On the other hand, in the context of VC, we propose to use the F-ratio for the analysis of features corresponding to the converted voices. In particular, the task in the VC is to map the source speakers' features to the target speakers' features as accurately as possible. Hence, the goal is to get *lower* values of F-ratio between a class corresponding to the converted voices (C_{vc}), and a class corresponding to the target speakers' voices (C_t) across all the available frequency bands. Here, we used the cepstral features instead of the spectral features to calculate the F-ratio. The F-ratio value is given by [167]:

$$F_i = \frac{(\mu_i^{vc} - \mu_i^t)^2}{\frac{1}{N_{vc}} \sum_{x_i \in C_{vc}} (x_i - \mu_i^{vc})^2 + \frac{1}{N_t} \sum_{x_i \in C_t} (x_i - \mu_i^t)^2}, \tag{4.7}$$

where x_i is the i^{th} cepstral coefficient of the MCC feature vector, $\mathbf{x}$. In addition, N_{vc} and N_t are the total numbers of frames in the class corresponding to the converted features, and the target speaker's spectral features, respectively.
The μ_i^{vc} and μ_i^t are the mean value of the x_i of all the frames corresponding to the converted voice, and the actual target speech, respectively. The F-ratio values for different frequency bands will form the F-ratio pattern, i.e., $[F_1, F_2, ..., F_d]$, where d is the dimension of the MCC features. Figure 4.10 shows the F-ratio analysis for the CMU-ARCTIC and the VCC 2016 database. Here, F-ratio is calculated between the class of converted voices and the actual target speakers' voices for all the systems developed using S1, S2, S3, and S4 methods. The lower the value of F-ratio, lesser the discrimination between features of the target speaker, and the converted voice, indicating that the proposed system is better in the context of the VC task.

F-ratio scores for the S1 are lower than the proposed techniques for some frequency bands (not for all). However, we can still see that the baseline system S1 is mostly having higher F-ratio than any one of the proposed outlier removal systems (i.e., S2, S3, and S4) for the feature dimension corresponding to the mid-to-high frequency bands. We consider 0-2667 Hz, 2667-5333 Hz, and 5333-8000 Hz as low, mid, and the high frequency bands, respectively. Moreover, these mid-to-high frequency bands are corresponding to the higher formants (i.e., third, fourth, and fifth formants) in the spectral features. The higher formants are well known to capture more speaker-specific information than the lower formants (i.e., first, and second) [168]. Hence, obtaining the lower F-ratio scores with the proposed methods in the mid-to-high frequency bands clearly indicates that the effectiveness of the proposed methods in transforming the speaker-specific information from the source to the target speaker. Thus, the effectiveness of outlier removal as the pre-processing can be seen in almost all the cases considered here for both the databases by observing the lower value of the F-ratio in at least any one of the systems S2, S3, and S4 compared to the baseline S1.



(a) GMM-based VC (D1)

(b) GMM-based VC (D2)

(c) ANN-based VC (D1)

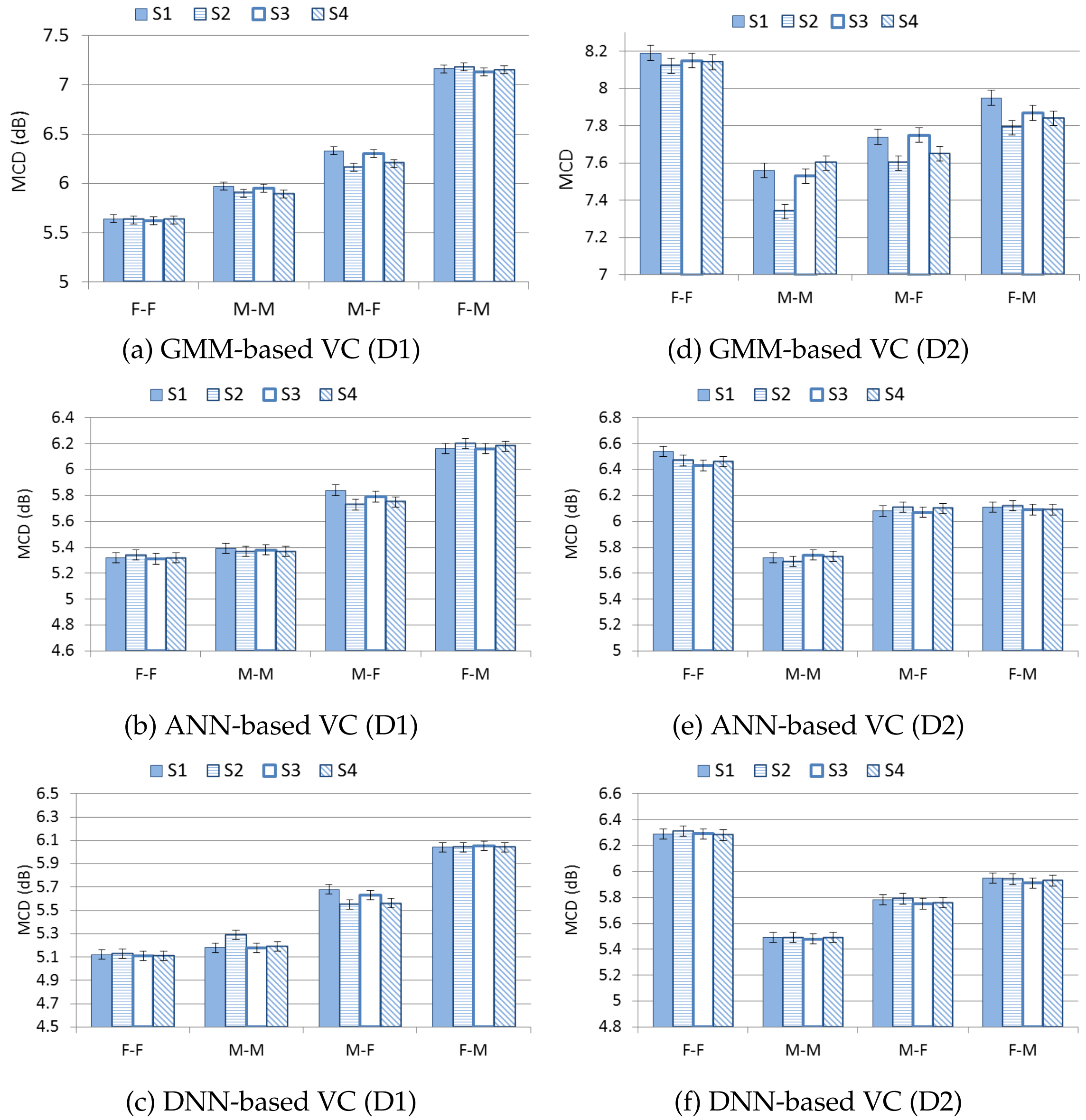
(d) ANN-based VC (D2)

(e) DNN-based VC (D1)

(f) DNN-based VC (D2)

Figure 4.10: F-ratio analysis for developed VC systems with different VC techniques: (a), (c), (e) for D1: CMU-ARCTIC database, and (b), (d), and (f) for D2: VCC 2016 database. S1 = baseline system, S2-S4 = proposed systems (see Table 2 for details). After [5].

From Figure 4.10, it can be observed that the F-ratio values are one order of magnitude smaller for the VCC 2016 database than for CMU-ARCTIC database. One of the possible reasons for this could be that the number of outliers detected in the case of VCC 2016 database is more compared to the CMU-ARCTIC database. In particular, we found on an average 8.38% and 6.88% of the total data are detected as outliers in the case of VCC 2016 and CMU-ARCTIC, respectively. Hence, detection and removal of more outliers in the case of VCC 2016 might lead to the better value of F-ratio. Furthermore, the amount of training data available in the VCC 2016 is less compared to the CMU-ARCTIC database. Hence, obtaining better performance in terms of F-ratio is an advantage, since all the real-world applications of VC suffer from the issue of less amount of available training data. In this thesis, the key idea is to observe the effectiveness of outlier removal in the context of VC by obtaining relatively lesser values of F-ratio for the converted voices using the proposed outlier removal method than the baseline method. However, the relation between absolute value of F-ratio, and subjective scores of the converted voices is yet to be studied in detail and is open research problem.



(a) GMM-based VC (D1)

(d) GMM-based VC (D2)

(b) ANN-based VC (D1)

(e) ANN-based VC (D2)

(c) DNN-based VC (D1)

(f) DNN-based VC (D2)

Figure 4.11: MCD analysis along with 95 % confidence interval for different speaker-pairs with different VC types: (a), (c), (e) for D1: CMU-ARCTIC database, and (b), (d), and (f) for D2: VCC 2016 database. After [5].

The traditional objective measure, MCD is considered here [16]. Figure 4.11 shows the details of average MCD analysis for the three different types of VC systems for various intergender, and intragender speaker-pairs along with a 95 % confidence interval for the statistical significance of the results. The 95 % confidence interval indicates that there is a 0.95 probability that the population mean (in this case, the MCD) will lie in that range (i.e., the margin of error) [169]. For statistically significant results, the margin of error should be as small as possible. We have shown the error bars corresponding to the 95 % confidence interval analysis in Figure 4.11. Furthermore, it is clear that for all the cases, at least one of outlier removal techniques is performing better than the baseline system S1. In particular, we have observed on an average 0.52% and 0.6% relative reduction in the MCD with the proposed outlier removal techniques compared to the S1 system for the CMU-ARCTIC and the VCC 2016 database, respectively. In addition, we have observed that there is relatively more reduction in the MCD for intergender VC systems compared to the intragender VC systems. It clearly indicates that the outlier removal is more significant in the intergender VC systems. These results indicate that there is a more chance of occurrence of outliers in the intergender VC task. In fact, for the CMU-ARCTIC database, we found on average 7.76% and 6% of the total data are detected as outliers for intergender and intragender pairs, respectively. Similarly, for the VCC 2016 database, on an average 8.75% and 8% of the total data are detected as outliers for intergender and intragender pairs, respectively. This shows that there is indeed more occurrence of outliers in the case of intergender systems compared to the intragender systems. In this context, such outliers may be due to the differences in size and shape of vocal tract for the source and the target speakers, glottal source characteristics (change in F_0, and its dynamics w.r.t. gender) and most important prosodic changes (such as, speaking style, which predominately indicates the manner in which articulators are used to produce an intelligible speech) in features due to the gender variability.

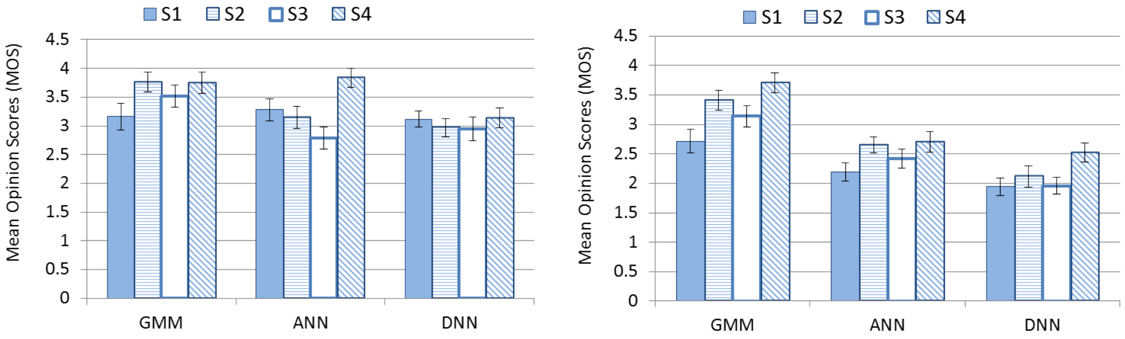
Moreover, there seem to be no statistically significant improvements in the MCD for DNN-based VC in the case of the CMU-ARCTIC database (as shown in Figure 4.11 (c)). It is possibly due to the fact that DNN-based VC techniques may capture more complex, and *nonlinear* relationships between the source and target speakers' spectral features compared to the traditional GMM-based systems. Similar observation was found for the VCC 2016 database. In fact, we can see that there is progressively less reduction in MCD in the performance of outlier removal techniques with ANN and DNN-based systems compared to the GMM-based systems. Hence, the ANN and DNN systems are less susceptible to the outliers compared to the GMM-based systems. However, this needs further investigations and it is an open research problem. Still the proposed pre-processing indeed helps to achieve the lower values of the MCD in ANN, DNN, and GMM-based VC systems.

4.4.3 Subjective Evaluation

Here, we have selected four speaker-pairs from both the databases (i.e., CMU-ARCTIC, and VCC 2016). These four speaker-pairs consist of two intragender (i.e., male-male, female-female), and two intergender (i.e., male-female and female-male) pairs for subjective evaluations. Mean Opinion Score (MOS) for evaluating the quality (i.e., the naturalness of the converted voice) has been used for the subjective evaluation. In addition, the comparative subjective test, namely, ABX test has been selected for evaluating the *speaker similarity* (SS) of the converted voice w.r.t. the actual target speaker's speech. High-quality headphones (Sennheiser HD280pro) were used for both the subjective tests. A total of 30 listeners (22 males, and 8 females with 18-35 years of age, and with no known hearing impairments) participated in both the subjective tests. Voice samples corresponding to the converted voices can be downloaded from the URL given in [170].

4.4.3.1 Mean Opinion Score (MOS) Test

In total, 48 converted voices were taken for the MOS evaluation from the VC systems developed on each database. Each randomly played utterance has been rated by the subjects for the *voice quality* (i.e., naturalness) on a five-point scale (1=Bad, 2=Poor, 3=Fair, 4=Good, 5=Excellent) [171].



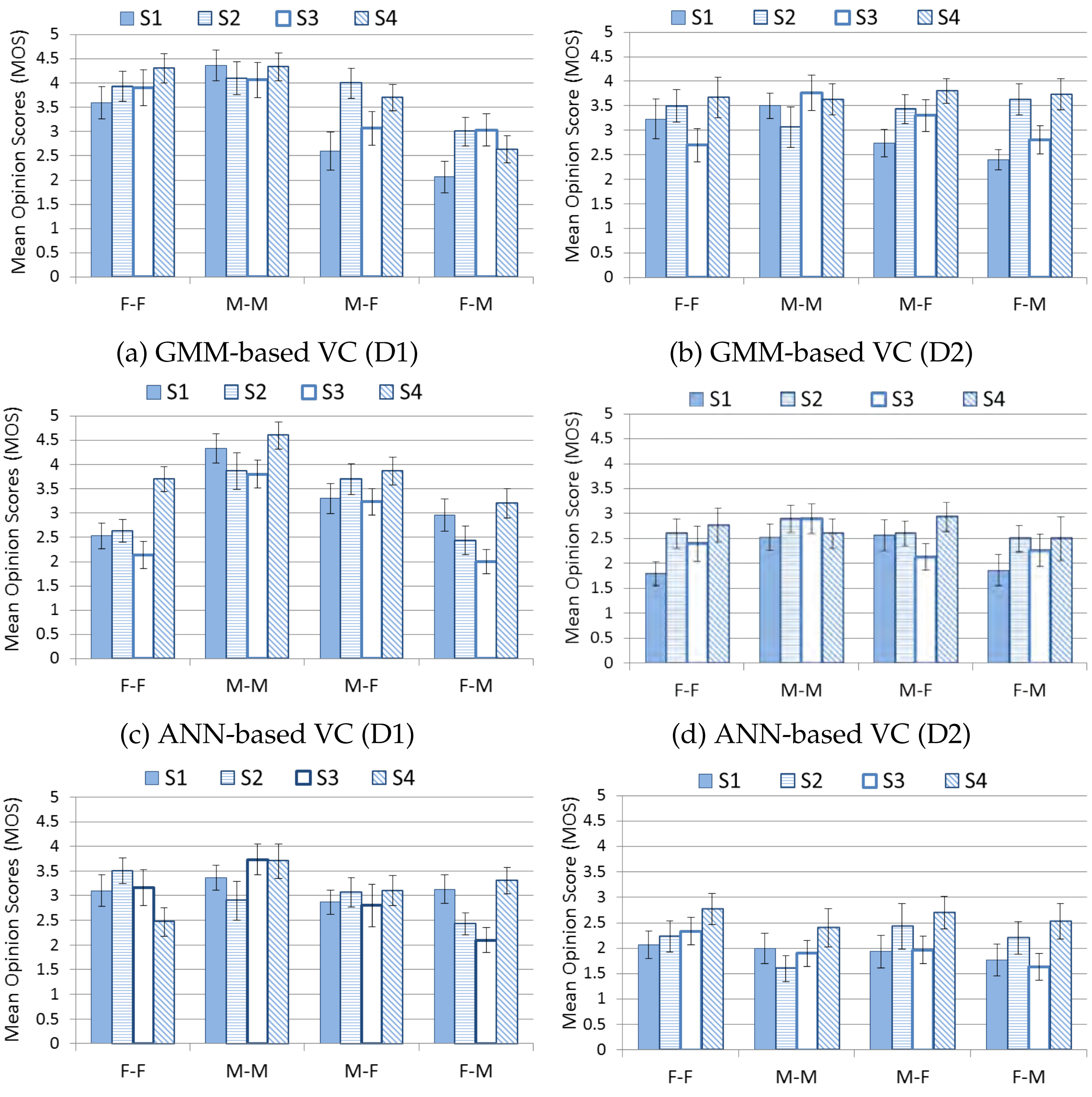
(a) D1

(b) D2

Figure 4.12: MOS scores using a 95 % confidence interval for GMM, ANN, and DNN-based VC systems for (a) D1: CMU-ARCTIC database, and (b) D2: VCC 2016 database. After [5].

Figure 4.12 shows the average MOS analysis along with the 95 % confidence interval for GMM, ANN, and DNN-based VC methods on both the databases. The smaller width of a margin of error in Figure 4.12 clearly demonstrates the statistical significance of the experimental results. It is clearly observed from Figure 4.12 that for both the databases, MOS is increasing with the proposed outlier removal techniques compared to the baseline system S1. In particular, there is a relative improvement of 3.34%, -3.14%, and 12.24% (on an average) in MOS for the CMU-ARCTIC database with the systems S2, S3, and S4 compared to the S1, respectively. Similarly, there is a relative improvement of 19.42%, 9.93%, and 30.51%

(on an average) in MOS for the VCC 2016 database with S2, S3, and S4 compared to the S1, respectively.



(a) GMM-based VC (D1) (b) GMM-based VC (D2)

(c) ANN-based VC (D1) (d) ANN-based VC (D2)

(e) DNN-based VC (D1) (f) DNN-based VC (D2)

Figure 4.13: MOS scores using a 95 % confidence interval for different speaker-pairs with different VC types, such as GMM, ANN, and DNN: (a), (c), (e) for D1: CMU-ARCTIC database, and (b), (d), and (f) for D2: VCC 2016 database. After [5].

This analysis clearly indicates that amongst the three different types of VC, S4 is performing better than the other two proposed outlier removal techniques. Since cut-off for the S3 is higher than the S2 and S4, the number of outliers detected and removed will be less in the case of S3 compared to the S2 and the S4. Hence, the improvement in terms of MOS in the case of the S3 is less compared to the other outlier removal methods. On the other hand, cut-off for the S2 is lower compared to the S4 because of which it may be removing some of the important pairs that are marked as outliers by the S2. Good data points appearing as the outliers are known as a *swamping* in the literature [156]. Hence, the phenomenon of swamping could be one of the reasons for getting slightly less improvement with the S2 compared to the S4.

Figure 4.13 shows the detailed analysis of the MOS for the intragender and the intergender cases for all the three different types of VC systems with 95 % confidence interval. Due to fewer training utterances available for a given speaker-pair in the VCC 2016 database compared to the CMU-ARCTIC database, it was observed that the MOS is lower for the systems developed on the VCC 2016 database compared to the systems developed on the CMU-ARCTIC database. Though MOS of the systems developed on the VCC 2016 database is less compared to the systems developed on the CMU-ARCTIC database, % relative improvement obtained after the outlier removal is more in the case of the VCC 2016 database. It may be due to the fact that the amount of available training data is more in the case of the CMU-ARCTIC database compared to the VCC 2016 database. Hence, the removal of the outliers perceptually (revealed via MOS) brings less improvement in terms of MOS for the CMU-ARCTIC database compared to the VCC 2016 database, which is in contrast with the MCD analysis. In the VC literature, other studies also reported that the MCD scores do not always correlate well with the subjective scores [12, 14, 43–45, 172].

Furthermore, we have observed that there is a relative improvement of 18.67%, 17.07% and 0.64% in MOS with the S4 system compared to the S1 for the GMM, ANN, and DNN-based systems, respectively. Similarly, for the VCC 2016 database, there is a relative improvement of 36.4%, 23.29%, and 30.41% in MOS with S4 system compared to the S1 for the GMM, ANN, and DNN-based systems, respectively. Hence, we can see that there is a relative reduction in the improvement of MOS with the proposed system S4 for the DNN, and ANN-based systems compared to the GMM-based systems. Similar to the MCD analysis, these results also indicate that the DNN and the ANN-based systems can handle the presence of outliers more robustly than the GMM-based systems while learning the mapping function. However, improvement in all the cases clearly indicates that the proposed pre-processing is still required. In addition, we have observed that there is relatively more improvement in the MOS for intergender VC systems compared to the intragender VC systems. It indicates that the outlier removal is more significant for the intergender VC systems as discussed in sub-Section 4.2.2.

4.4.3.2 ABX test

We found that the S4 is performing relatively best among all the proposed techniques w.r.t. both the MCD and MOS tests. Hence, ABX tests have been conducted for measuring the speaker similarity (SS) between the baseline system S1 and the S4. In this test, the same utterances converted using two different approaches were played randomly. Subjects were asked to choose the utterances that were closer to the target speaker in the context of SS. Figure 4.14 shows the analysis of ABX tests.

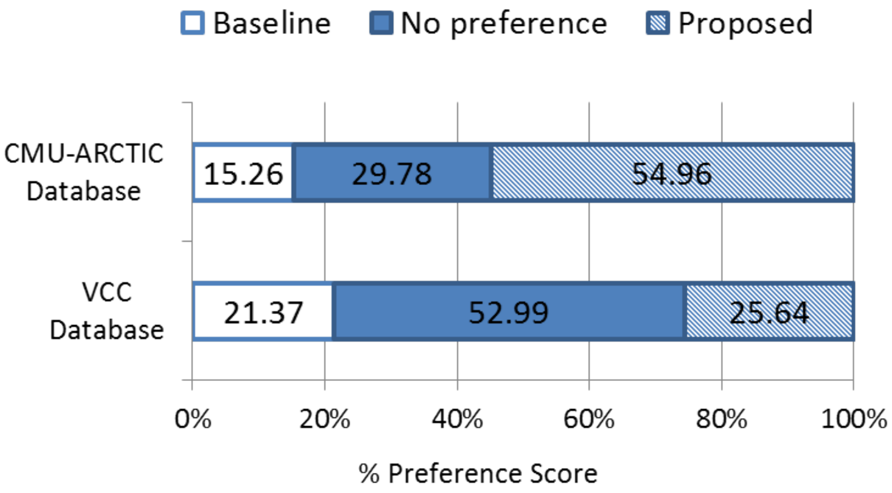


Figure 4.14: ABX test result for SS. Here, the margin of error corresponding to the 95 % confidence interval is 0.06 for both the databases. After [5].

For the ABX tests, total nine samples have been taken from the systems developed using GMM, ANN, and DNN-based methods. It is visible from Figure 4.14 that the proposed system S4 is absolutely 39.7% and 4.27% more preferred than the S1 for the VC systems developed on the CMU-ARCTIC, and VCC 2016 databases, respectively. However, we have observed that for 29.78% and 52.99% of the times subjects were not able to discriminate which systems were better in terms of SS for the CMU-ARCTIC, and VCC 2016 databases, respectively. In VC techniques based on ML optimization, such as GMM, ANN, and DNN, the objective function of the training is to minimize the numerical error between the converted and the actual target speakers' spectral features. Hence, during the training, parameters of the network are adjusted in such a way that the numerical error is minimized. Hence, the objective scores are important since the goal of training is to minimize the error. However, the reduction in numerical estimates does not always correlate well with the generated sample quality. Moreover, the ML-based optimization of the network parameters does not always lead to the better perceptual speech quality. Hence, more importance has been given to the subjective scores compared to the objective score in the VC literature. Furthermore, it is very difficult (if not impossible) to capture the challenges of sophisticated human perception for hearing via these kinds of naive objective measures. Still, the training of VC itself focuses on the objective scores. Hence, the better training will lead to the improved objective scores. Thus, we believe that it is desirable to obtain better performance in both objective and subjective scores in the context of the VC task.

4.5 Proposed Outlier Removal Approach *vs.* Robust Alignment Strategies

In this sub-Section, we compare the performance of the proposed outlier removal technique (S4) against the robust alignment strategies. In particular, one of the reasons for poor alignment is the occurrence of speech-silence pairs (i.e., speech *vs.* non-speech pairs). It is true that one can use Voice Activity Detection (VAD) to remove the silence frames of an utterance (for both source and the target speakers), however, as mentioned in Section 2.2, apart from these silence or non-speech regions, non-linear warping of DTW, and the spectral feature variations across the speakers (especially across genders, such as female, and male speakers) are also sources of outliers in the context of VC. In addition, the VAD may sometimes detect low energy unvoiced regions as the non-speech especially in the case of fricatives. Hence, it may remove those regions as well in addition to the non-speech regions. Hence, we did not use VAD earlier and in fact, we propose to use outlier removal as an independent technique that could remove all such pairs that do not follow the general trend of the data before learning the mapping function.

Table 4.2: MCD analysis of various systems developed. After [5]

	CMU-ARCTIC			**VCC 2016**		
	S1	**VAD**	**S4**	**S1**	**VAD**	**S4**
GMM	6.13	6.23	**6.12**	7.1	7.17	**7.085**
ANN	5.46	5.56	**5.43**	5.66	5.82	**5.66**
DNN	5.51	5.65	**5.48**	5.8	5.98	**5.8**

In this Section, we show preliminary results using VAD. We have applied VAD instead of the outlier removal technique for the training of different VC systems.

We used a COoperative Voice Analysis REPository for speech technologies (COVAREP) 1.4.1 toolbox for VAD [173]. This technique uses combined source and filter-based robust features, which are further used to train the ANN classifier using multi-conditioned dataset [174]. The MCD results for the developed various systems are shown in Table 4.2. We have compared the results with the baseline (S1), and the proposed outlier removal algorithm (S4). It can be clearly seen that the systems that are developed using VAD pre-processing have obtained an average 2.15% relative increment in the MCD compared to the baseline, which is undesirable. The key reason for this could be the removal of silence-silence in addition to the silence-speech pairs from the training due to VAD. In this thesis, we believed that these undesirable silence-speech pairs can be captured using outlier removal-based techniques instead of VAD. In fact, we found out that on an average 79.22% and 88.57% of the total detected outliers are correspond to the silence-speech pairs for CMU-ARCTIC and VCC 2016 databases, respectively.

Furthermore, differences between the source and the target speakers' spectral representations, which is primarily due to differences in the vocal tract system (shape and size) could be one of the reasons for the generation of outliers in the context of VC. To alleviate this, the Vocal Tract Length Normalization (VTLN) technique could be useful to generate a more robust correspondence between source and the target speakers' spectral representations. To measure the effectiveness of the proposed outlier removal approach over the VTLN-based techniques, we have developed VC systems using the VTLN-based alignment technique on both CMU-ARCTIC and VCC 2016 databases. In particular, an all-pass transform based on a bilinear function is applied to perform VTLN. This technique was originally proposed for a speech recognition task [175], and its implementation in the cepstral-domain was discussed in the [176]. This VTLN technique is one of the state-of-the-art techniques in the VC literature [22, 42, 89].

Table 4.3: MCD analysis of various systems developed. After [5]

	CMU-ARCTIC			**VCC 2016**		
	S1	**VTLN**	**S4**	**S1**	**VTLN**	**S4**
GMM	6.13	**6.05**	6.12	7.1	**7.04**	7.085
ANN	5.46	5.46	**5.43**	5.66	5.73	**5.66**
DNN	5.51	5.58	**5.48**	5.8	5.8	**5.8**

The MCD analysis of the developed VC systems is shown in Table 4.3. We obtained on average -0.06% relative reduction in the MCD compared to the baseline S1. On the other hand, with the proposed method (i.e., S4), we have obtained an average 0.25% relative reduction in the MCD compared to the baseline S1. Furthermore, we observed that more computational time (approximately by a factor of 10.7 [13]) is required to obtain the aligned feature pairs using the VTLN technique than the proposed outlier removal approach. Training times for all the approaches are reported in Table III. Experiments are performed on a DELL I7 machine with 16 GB RAM, and a 800 MHz clock cycle. Hence, the effectiveness of our proposed outlier removal approach over VTLN-based alignment can be easily seen from Table 4.4.

Table 4.4: Computation time required (in seconds) to obtain aligned pairs using different approaches. After [5]

	S1	VTLN	S4
CMU-ARCTIC	86.54 s	976.35 s	88.21 s
VCC 2016	35.19 s	357.87 s	36.44 s

4.6 Chapter Summary

In this chapter, we presented the DTW algorithm, which is widely used alignment strategy for parallel VC. In particular, limitations of the DTW algorithm in the context of VC is presented. After alignment, some corresponding feature pairs are still inconsistent with the rest of the data and are considered as outliers. These outliers shift the parameters of mapping function from their true value and hence, negatively affect the learning of mapping function during the training phase of the VC task. In this chapter, we proposed the novel pre-processing technique of outlier removal before learning the mapping function in various state-of-the-art technique for parallel VC.

We identified one of the possible causes of outliers in the context of the VC. We have presented the proposed outlier removal technique in detail. The proposed method uses the score distance that is estimated using the ROBPCA to detect the outliers. In particular, the outliers are determined using a fixed cut-off based on the degrees of freedom in a chi-squared distribution, which is speaker-pair independent. One of the reasons for selecting the fixed cut-off is the assumption that the score distances follow the normal distribution. However, we found that the score distances do not always follow the normal distribution in the context of the

VC task. Hence, we propose to explore the speaker-pair-dependent cut-offs to detect the outliers. In particular, we found the best cut-off value is the $\mu + 2\sigma$, calculated on the logarithm of the score distances. We selected the state-of-the-art GMM, ANN, and DNN-based VC systems to evaluate the effectiveness of the outlier removal in the context of the VC. We have presented our results on two state-of-the-art databases, namely, CMU-ARCTIC, and VCC 2016.

Furthermore, we have presented a detailed analysis of objective and subjective evaluations using a 95 % confidence interval. We proposed to use the F-ratio-based analysis for objective evaluations of the VC system. We found a better performance of the proposed systems compared to the baseline VC systems using the F-ratio-based analysis. We have also performed the MCD test for objective evaluation. In particular, we obtained an average 0.56% relative decrease in MCD with the proposed outlier removal technique as a pre-processing step. In particular, we have observed that there is relatively more reduction in the MCD for intergender VC systems compared to the intragender VC systems. Furthermore, for large training data, the task of outlier removal is *not* that significant. However, in the context of the VC task, the large amount of training data is not available from the target speaker in most of the real-world applications. Hence, the task of outlier removal becomes essential in real-world applications of VC.

Furthermore, improvements in the speech quality, and the speaker similarity were observed with the proposed speaker-pair-dependent cut-off based outlier removal. From both the subjective and objective evaluations, we found that the effectiveness of the outlier removal is more in the GMM-based VC systems compared to the ANN and DNN-based VC systems. This observation indicates that the ANN and DNN-based systems are less susceptible to the outliers than the GMM-based VC systems. The possible reason behind this could be the ability of the ANN and DNN to capture more complex, and nonlinear relationships between the source, and the target speakers' spectral features. In the next chapter, we will present different proposed alignment strategies for non-parallel VC.

CHAPTER 5

Alignment Strategies for Non-parallel VC

5.1 Introduction

In chapter 4, we described the alignment strategies, and its impact on the quality of converted voices for the parallel VC task. In addition, we proposed outliers removal strategies as a pre-processing step to tackle the issues related to the alignment in parallel VC. In this chapter, we present a few novel alignment strategies for non-parallel VC. The proposed convergence theorem for **I**terative combination of a **N**earest Neighbor search step and a **C**onversion step **A**lignment (INCA) algorithm is discussed in Section 5.2. Effectiveness of the dynamic features for calculating the Nearest Neighbor (NN) search path in the INCA, and Temporal Context (TC)-INCA is described in Section 5.3. Section 5.4 presents the NN-based alignment strategy, which uses the metric learning technique for finding the NN search path. Furthermore, phone aware NN-based alignment technique that uses Spectral Transition Measure (STM) algorithm for estimating the phonetic boundaries is discussed in Section 5.5, whereas Section 5.6 concludes the chapter.

5.2 INCA Algorithm and its Convergence

5.2.1 Iterative combination of a Nearest Neighbor search step and a Conversion step Alignment (INCA) Algorithm

INCA algorithm iteratively performs three steps, namely, a nearest neighbor search step, training of mapping function and transformation step using JDGMM-based VC until convergence. Graphical representation of INCA algorithm is shown in Figure 5.1. Let $\mathbf{X} = \{\mathbf{x}_k\}_{k=1}^{N_x}$, $\mathbf{Y} = \{\mathbf{y}_j\}_{j=1}^{N_y} \in \mathbb{R}^d$ be the spectral features obtained from the non-parallel corpus of the source and target speakers, respectively. The alignment procedure is given below using *asymmetric-1* variant of INCA algorithm as it is considered the best among all other variants of INCA algorithm [6].

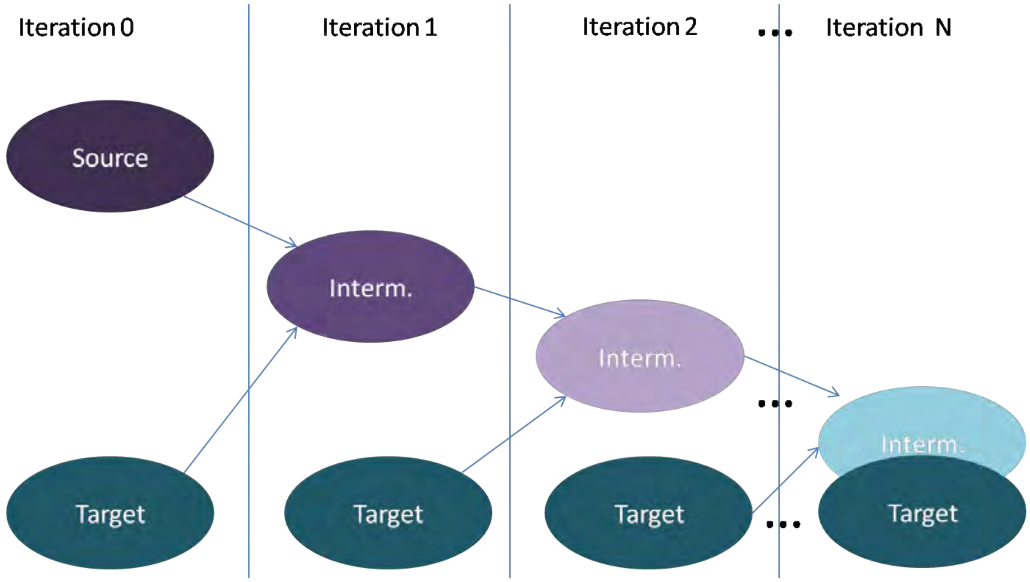


Figure 5.1: Graphical representation of INCA algorithm. After [6].

However, it can be easily extended to the other variants of INCA algorithm.

1. **Initialization:** At t^{th} iteration, auxiliary vector set, i.e., $(\mathcal{F}_{t-1}(\{\mathbf{x}_k\}) = \{\mathbf{x}_k^{'}\})$ represents an intermediate acoustic space of converted spectral features of previous iteration. The mapping function is initialized as $\mathcal{F}_0(\mathbf{x}) = \{\mathbf{x}_k\}$, which is also called trivial initialization.

2. **NN search:** At each iteration for each vector $\mathbf{x}_k^{'}$, the index of its corresponding NN vector in $\mathbf{Y}$ is estimated and stored in p(k). Similarly, for each vector $\mathbf{y}_j$, its corresponding NN vector is found from $\{\mathbf{x}_k^{'}\}$ and stored its index in $q(j)$.

$$p_t(k) = \arg\min_j d(\mathcal{F}_{t-1}(\mathbf{x}_k), \mathbf{y}_j), \quad q_t(j) = \arg\min_k d(\mathbf{y}_j, \mathcal{F}_{t-1}(\mathbf{x}_k)), \tag{5.1}$$

   where $d(.)$ is the Euclidean distance.

3. **Training:** The spectral feature vectors given by $\{\mathbf{x}_k, \mathbf{y}_{p(k)}\}$ and $\{\mathbf{x}_{q(j)}, \mathbf{y}_j\}$ are concatenated, trained, and then mapping function $\mathcal{F}_t(.)$ is applied using JDGMM-based method [19].

4. **Transformation:** The auxiliary vector set $\mathbf{X}^{'}$ is updated after applying the mapping function $\mathcal{F}_t()$, i.e.,

$$\mathbf{x}_k^{'} = \mathcal{F}_t(\mathbf{x}_k), \forall k. \tag{5.2}$$

5. **Convergence Checking:** If the converted spectral features are very near to the target spectral features in mean square error (MSE) sense then convergence is achieved, otherwise go to the *step 2*. The MSE between intermediate converted vectors, and the target vectors is given by:

$$E_T = \frac{1}{N_x + N_y}\left(\sum_{k=1}^{N_x} \|\mathcal{F}_t(\mathbf{x}_k) - \mathbf{y}_{p_t(k)}\|^2 + \sum_{j=1}^{N_y} \|\mathbf{y}_j - \mathcal{F}_t(\mathbf{x}_k)_{q_t(j)}\|^2\right), \tag{5.3}$$

where $||.||^2 = \sum_{<n>} |x(n)|^2$ (i.e., square of l^2 norm). The empirically convergence of E_t was shown in [6].

Table 5.1: Variants of INCA algorithm. Adapted from [6]

Name	INCA algorithm: Step 2	
Asymmetric-1	$p_t(k) = \arg\min_j d(\mathcal{F}_{t-1}(\mathbf{x}_k), \mathbf{y}_j)$,	$q_t(j) = \arg\min_k d(\mathbf{y}_j, \mathcal{F}_{t-1}(\mathbf{x}_k))$
Asymmetric-2	$p_t(k) = \arg\min_j d(\mathbf{x}_k, \mathcal{F}_{t-1}^{-1}(\mathbf{y}_j))$,	$q_t(j) = \arg\min_k d(\mathcal{F}_{t-1}^{-1}(\mathbf{y}_j), \mathbf{x}_k)$
Symmetric-1	$p_t(k) = \arg\min_j d(\mathcal{F}_{t-1}(\mathbf{x}_k), \mathbf{y}_j)$,	$q_t(j) = \arg\min_k d(\mathcal{F}_{t-1}^{-1}(\mathbf{y}_j), \mathbf{x}_k)$
Symmetric-2	$p_t(k) = \arg\min_j d(\mathbf{x}_k, \mathcal{F}_{t-1}^{-1}(\mathbf{y}_j))$,	$q_t(j) = \arg\min_k d(\mathbf{y}_j, \mathcal{F}_{t-1}(\mathbf{x}_k))$

NN pairs in the second step of the INCA can be obtained using three other ways (as defined in Table 5.1). The different variants of the INCA use different combinations of the feature set to calculate NN pairs. They are primarily classified in two categories, namely, asymmetric and symmetric. In asymmetric variants, alignment results will be changed if source and the target speakers are interchanged. On the other hand, alignment results will not be changed in the symmetric variants if both the speakers are interchanged. Here, the mapping function $\mathcal{F}(.)$ is given by the JDGMM-based transformation, which is defined in Eq. (3.2) in Chapter 3. Moreover, $\mathcal{F}^{-1}(.)$ is the inverse mapping function, i.e., $\mathcal{F}^{-1}(\mathbf{Y}_t) = E(\mathbf{X}_t|\mathbf{Y}_T)$ based on MMSE criteria is given by:

$$\hat{\mathbf{x}} = \mathcal{F}^{-1}(\mathbf{y}) = \sum_{m=1}^{N_c} p_m(\mathbf{y})(\mu_m^{(\mathbf{x})} + \Sigma_m^{(\mathbf{xy})}(\Sigma_m^{(\mathbf{yy})})^{-1}(\mathbf{y} - \mu_m^{(y)})), \tag{5.4}$$

where $p_m(\mathbf{y}_t) = P(m|\mathbf{y}_t, \lambda^{\mathbf{z}}) = \frac{\omega_m \mathcal{N}(\mathbf{y}|\mu_m^{\mathbf{y}}, \Sigma_m^{\mathbf{yy}}))}{\sum_{k=1}^{M} \omega_k \mathcal{N}(\mathbf{y}|\mu_k^{\mathbf{y}}, \Sigma_k^{\mathbf{yy}}))}$ is the posterior probability of the target vector, $\mathbf{y}_t$, for the m^{th} Gaussian component, which is similar to defined in Section 3.2 in Chapter 3. All the variants of the INCA are valid. However, it has been shown that the asymmetric-1 variant of the INCA performs empirically better compared to the other variants [6]. In addition, step-2 of the asymmetric-1 variant of INCA is straightforward in terms of mapping source to the target. On the other hand, other variants of INCA require to compute inverse mapping function, which is against the philosophy of source to the target conversion. Hence, we presented our analysis on asymmetric-1 variant of the INCA. However, our analysis can be easily extended to other variants of the INCA.

The INCA iteratively aligns all the training vectors in **X** and **Y**. Initially, there may be some source feature vectors that do not have any target feature vector in their neighborhood, this will generate one-to-many aligned pairs. The converted feature vectors that are obtained from the mapping function trained on such one-to-many aligned pairs move these converted feature vectors gradually to the acoustic space of the target feature vectors. As the number of iterations increase, the alignment becomes more and more accurate. Here, any mapping function (discussed in Chapter 3) can be applied. However, there are primarily two key reasons for selecting the JDGMM-based mapping function. For example, the JDGMM-based mapping function uses conditional expectation, which is the best MMSE operator, and hence, it will be helpful in developing the convergence theorem for the INCA that is discussed in the next sub-section. In addition, the oversmoothing issue presents in the JDGMM-based method will help to minimize the effect of one-to-many and misaligned feature vectors in the INCA. Though over-smoothing is reported to produce a noticeable decrease in the quality of the converted voices, it is advantageous in the INCA, since it minimizes the alignment errors. Once the alignment is finished, any mapping function can be applied to learn the mapping for the non-parallel VC task.

5.2.2 Proposed Convergence Theorem

Following is the proposed proof of convergence for the INCA algorithm.

Theorem 1 *The INCA algorithm converges monotonically to a local minimum in MSE sense.*

Proof: At iteration t, in step 2 (i.e., eq. (5.1)), NN search path, i.e., $p_t(j)$ and $q_t(k)$ are calculated from the auxiliary set which is $\mathbf{X}' = \mathcal{F}_{t-1}(\mathbf{X})$, and target spectral feature set **Y**. The MSE between the correspondence is given by:

$$D_t = \left(\frac{1}{N_x + N_y}\right)\left(\sum_{k=1}^{N_x} ||\mathcal{F}_{t-1}(\mathbf{x}_k) - \mathbf{y}_{p_t(k)}||^2 + \sum_{j=1}^{N_y} ||\mathbf{y}_j - \mathcal{F}_{t-1}(\mathbf{x}_k)_{q_t(j)}||^2\right). \quad (5.5)$$

After that for given correspondence, $\mathcal{F}()$ is applied and the error between the converted spectral features and the target spectral features is given by:

$$E_t = \left(\frac{1}{N_x + N_y}\right)\left(\sum_{k=1}^{N_x} ||\mathcal{F}_t(\mathbf{x}_k) - \mathbf{y}_{p_t(k)}||^2 + \sum_{j=1}^{N_y} ||\mathbf{y}_j - \mathcal{F}_t(\mathbf{x}_k)_{q_t(j)}||^2\right). \tag{5.6}$$

As the mapping function is learned using Minimum Mean Square Error (MMSE) criteria, for every iteration, $E_t \leq D_t$. Now, after transformation for the next iteration, again NN search step is applied, and path $p_{t+1}(j)$ and $q_{t+1}(k)$ are learned. Hence,

$$\begin{aligned} ||\mathcal{F}_t(\mathbf{x}_k) - y_{p_{t+1}(k)}||^2 + ||\mathbf{y}_j - \mathcal{F}_t(\mathbf{x}_k)_{q_{t+1}(j)}||^2 \leq \\ ||\mathcal{F}_t(\mathbf{x}_k) - y_{p_t(k)}||^2 + ||\mathbf{y}_j - \mathcal{F}_t(\mathbf{x}_k)_{q_t(j)}||^2, \end{aligned} \tag{5.7}$$

i.e., $D_{t+1} \leq E_t$. Hence, mean square errors, namely, E_t and D_t must satisfy following inequality. In addition, MSE cannot be negative, and hence, lower bound occurs. In particular,

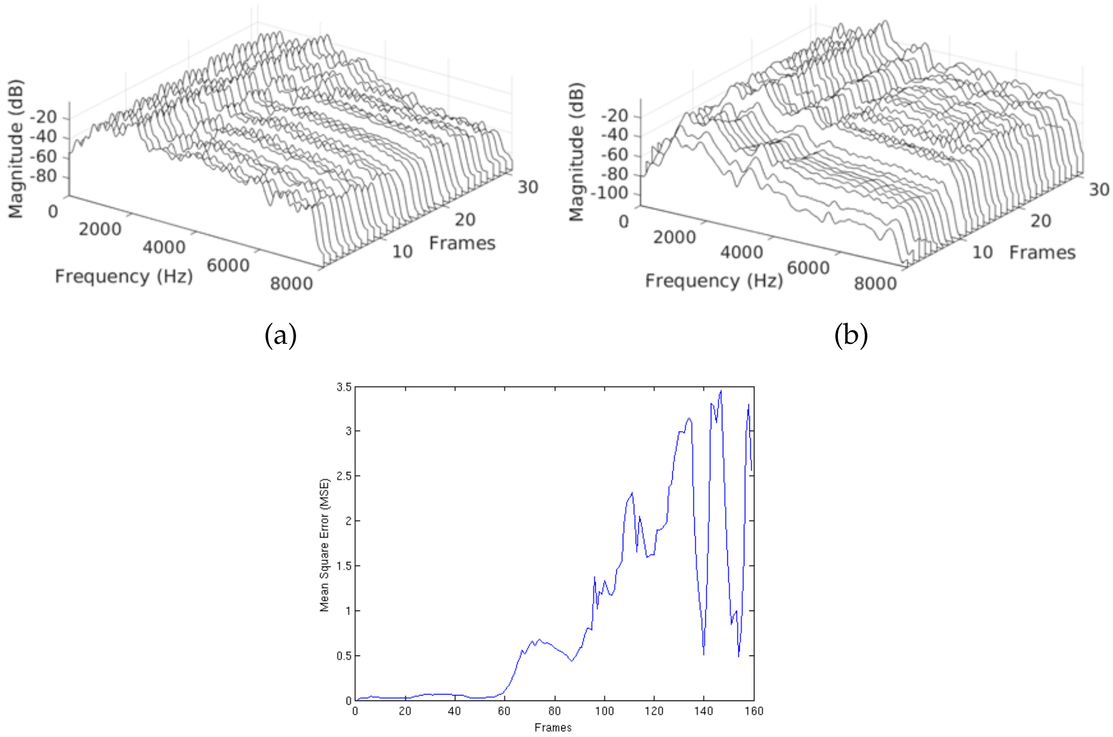
$$0 < E_{t+1} \leq D_{t+1} \leq E_t \leq D_t, \forall t. \tag{5.8}$$

Because the MSE sequence is non-increasing and bounded below, the subsequence E_t must converge monotonically as per the Bolzano-Weierstrass theorem, which states that every bounded sequence has a convergent subsequence [177]. ■

Hence, the key idea behind the convergence of the INCA algorithm is that MMSE-based conversion function used in GMM-based VC will decrease the average distance between the corresponding source and the target spectral features, whereas the nearest neighbor (NN) search will further reduce the distance for each spectral feature separately. This proof can also be extended to other three variants of the INCA algorithm (namely, asymmetric-II, symmetric-I, and symmetric-II) as the formulation of them requires the calculation of $\mathcal{F}^{-1}(.)$ that is also derived using MMSE criteria as discussed in Section 5.2.1.

5.2.3 Why Convergence in MSE Sense than Pointwise Convergence?

In VC task, $\mathbf{X}$ and $\mathbf{Y}$ are generally, spectral features, namely, Mel frequency cepstral coefficents (MFCC) of a source and target speakers, respectively. Computation of MFCC includes Mel frequency warping of spectral energy in cepstral-domain, and moreover, physical systems responds to signal's energy. Convergence in MSE sense will ensure that the source speaker's cepstral representation converges to the target speaker's cepstral representation. In particular, at the k^{th} frame, $\sum_i |e_k(i)|^2$, where $e_k(i) = \mathbf{x}_k(i) - \mathbf{y}_k(i)$, and i is the quefrency index of cepstrum. It means that $\mathbf{x}_k(i)$ and $\mathbf{y}_k(i)$ need *not* be equal at every value of 'i' rather there is *no* energy in their difference. On the other hand, pointwise convergence is more stricter than the MSE convergence. In particular, pointwise convergence implies MSE convergence, however, converse is not always true [178].



(a)

(b)

(c)

Figure 5.2: Waterfall plot of spectrum of a female twin-pair (both at the age of 27 years) uttering Hindi word /achanak/ (i.e., "Suddenly") (a) source, (b) target speaker, and (c) Mean Square Error (MSE). After [7,20]. Database of twin is taken from [25].

In the context of VC, pointwise convergence means that the cepstral representations of the converted cepstrum is exactly same to the target speaker's cepstrum at every instant of quefrency index, i. This is primarily due to the fact that one cannot have identical spectral representation and so cepstral representations due to physiological differences (related to vocal tract shape and size of source and target speaker). Thus, source speaker will never be able to exactly match the spectral representation of target speaker's vocal tract shape which in fact is absurd. However, one can minimize the difference between spectral representations w.r.t.

target speakers. Hence, convergence in MSE makes more sense in the context of VC. In this context, we have taken the same utterance from two twins speakers (i.e., a twin speaker-pair) who are having almost identical speaker characteristics (as they look and sound STMperceptually very similar) [179]. It can be seen from Figure 5.2 that even if they sound very similar, their time-varying spectral representations are very different and hence, the MSE between their spectrum is not zero at every point as shown in Figure 5.2 (c).

5.3 Proposed Dynamic Features for INCA and TC-INCA

Speech signal consists of various basic speech sound units, which are called as *phonemes*. These sounds and their features differ both in time and spectral characteristics [53]. The dynamic features, such as the change of distribution of spectral energy and temporal characteristics will play a vital role to discriminate major phonetic categories, such as nasals, stops, vowels, fricatives, etc. [180,181]. In this thesis, we propose to use the dynamic features to capture the contextual information that is present across the frames by taking the longer contextual frames as given in [78]. The dynamic feature is given by [181]:

$$\Delta \mathbf{x}_k = \frac{\sum_{i=1}^{T/2} i(\mathbf{x}_{k+i} - \mathbf{x}_{k-i})}{2\sum_{i=1}^{T/2} i^2}, \tag{5.9}$$

where T is even and contextual window length, $W_d = T + 1$, which is taken at the current frame, i.e., $\mathbf{x}_k$ by considering $T/2$ frames from the left and right side. Dynamic features are calculated and concatenated along with the static features. Hence, the new set of feature vectors is defined as $\mathbf{X}_k = [\mathbf{x}_k^T, \Delta\mathbf{x}_k^T]^T$, and $\mathbf{Y}_k = [\mathbf{y}_k^T, \Delta\mathbf{y}_k^T]^T$. Let $\mathbf{X} = \{\mathbf{X}_k\}_{k=1}^{\hat{N}_x}$, $\mathbf{Y} = \{\mathbf{Y}_j\}_{j=1}^{ST\hat{M}N_y} \in \mathbb{R}^d$, where $\hat{N}_x \leq N_x$, $\hat{N}_y \leq N_y$ are the number of feature vectors from the source and the target speakers, respectively. Here, the cost function which is given by eq. (5.3) is modified from [6,78] for our new set of feature vectors, and it is given by [7]:

$$\mathcal{E}_t = \frac{1}{\hat{N}_x + \hat{N}_y}\left(\sum_{k=1}^{\hat{N}_x} \|\mathcal{F}_t(\mathbf{X}_k) - \mathbf{Y}_{p_t(k)}\|^2 + \sum_{j=1}^{\hat{N}_y} \|Y_j - \mathcal{F}_t(\mathbf{X}_k)_{q_t(j)}\|^2\right), \tag{5.10}$$

where $\mathcal{F}(\mathbf{X}_k)$ is the transformation function [19] and the JDGMM is trained for joint feature vectors, which is obtained by concatenating the static and the dynamic features from both the source and target speakers. The cost function given by eq. (5.10) cannot be solved using the gradient descent algorithm due to its dependency on the mapping function, i.e., $\mathcal{F}(.)$ [182]. In such a scenario, alternating minimization technique is used. It is well known optimization technique that iteratively minimizes the cost function depending on more than one variables [183, 184]. Hence, dynamic INCA algorithm can be defined as an optimization problem, aiming to minimize the following cost:

$$\{p^*, q^*, \mathcal{F}^*\} = \underset{\{\mathcal{F},p,q\}}{\arg\min}\, \mathcal{E}(p, q, \mathcal{F}), \tag{5.11}$$

where p, q are the warping paths obtained after NN step 2 in the INCA. This joint optimization can be split into two separate minimization problems, which will be solved iteratively for $t = 1, 2, ...$ Hence, at iteration t, the algorithm (i.e., dynamic features in INCA (D-INCA) algorithm) is given by [7]:

$$\{p_t, q_t\} = \underset{\{p,q\}}{\arg\min}\, \mathcal{E}(p, q, \mathcal{F}_{t-1}), \tag{5.12}$$

$$\mathcal{F}_t = \underset{\mathcal{F}}{\arg\min}\, \mathcal{E}(p_t, q_t, \mathcal{F}). \tag{5.13}$$

5.3.1 Convergence of Proposed Dynamic INCA Algorithm

Due to alternating minimization nature of our training approach which is given by eq. (5.11) and eq. (5.12), it is clearly seen that the following inequality will always hold:

$$\mathcal{E}_t = \mathcal{E}(p_t, q_t, \mathcal{F}_t) \leq \mathcal{E}(p_t, q_t, \mathcal{F}_{t-1}) \leq \mathcal{E}(p_{t-1}, q_{t-1}, \mathcal{F}_{t-1}) = \mathcal{E}_{t-1}, \forall t. \tag{5.14}$$

Here, the cost function $\mathcal{E}_t$ is nothing but the Mean Square Error (MSE), and the transformation function is also given by the MMSE criteria as given in [19]. Hence, the above mentioned inequality will be non-increasing and bounded below, the subsequence $\mathcal{E}_t$ must converge monotonically as per the Bolzano-Weierstrass theorem [177]. However, convergence to the global minimum is not guaranteed as our cost function is *nonconvex*. Hence, it will converge to a local minimum. In particular, convergence is in MSE sense than the pointwise convergence as discussed earlier in Section 5.2.2.

5.3.2 Proposed Dynamic TC-INCA Algorithm

Similar to D-INCA, we propose to extend the idea of using dynamic features in the TC-INCA. The TC-INCA tries to use the Temporal Context information for finding NN-aligned pairs in the INCA algorithm [78]. In particular, this is achieved by concatenating the current spectral feature vector with the (T/2) successive feature vectors in both the sides, i.e., $\mathbf{X}_k = \{\mathbf{x}_{k-T/2}^T, ..., \mathbf{x}_k, ..., \mathbf{x}_{k+T/2}^T\}$, $\mathbf{Y}_k = \{\mathbf{y}_{k-T/2}^T, ..., \mathbf{y}_k, ..., \mathbf{y}_{k+T/2}^T\}$ for a given contextual window length (i.e., $W_c = T + 1$).

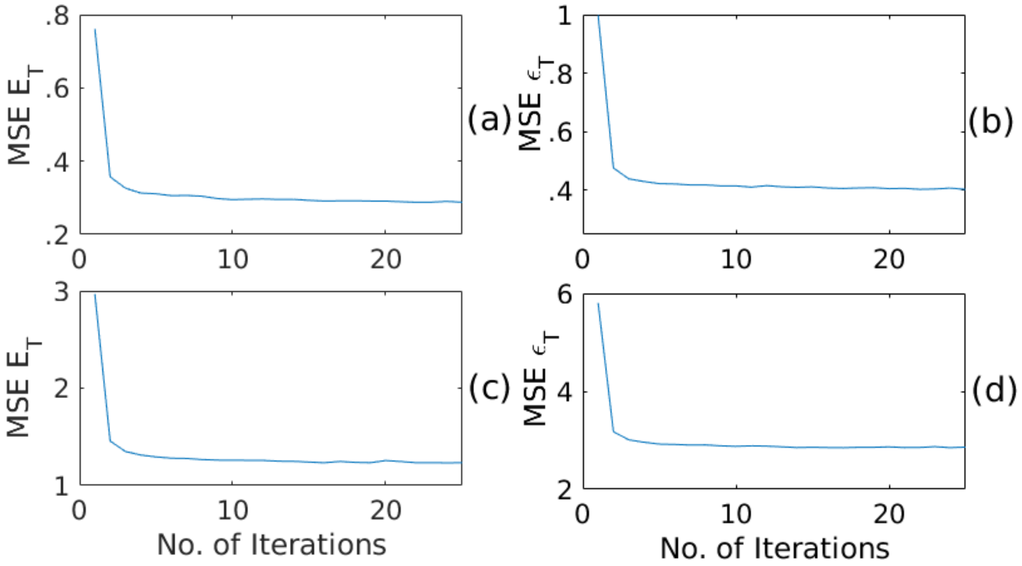


Figure 5.3: Empirical convergence analysis for (a) INCA, (b) D-INCA, (c) TC-INCA, and (d) D-TC-INCA. After [7].

The details of TC-INCA are given in [78]. Similarly, the spectral features will be considered along with its dynamic features, and the contextual features for finding the NN feature pairs for the proposed D-TC-INCA. In particular, the cost function given by Eq. (5.10) is the same except,
$\mathbf{X}_k = [\mathbf{x}_{k-T/2}^T, \Delta\mathbf{x}_{k-T/2}^T, ..., \mathbf{x}_k^T, \Delta\mathbf{x}_k^T, ..., \mathbf{x}_{k+T/2}^T, \Delta\mathbf{x}_{k+T/2}^T]^T$,
$\mathbf{Y}_k = [\mathbf{y}_{k-T/2}^T, \Delta\mathbf{y}_{k-T/2}^T, ..., \mathbf{y}_k^T, \Delta\mathbf{y}_k^T, ..., \mathbf{y}_{k+T/2}^T, \Delta\mathbf{y}_{k+T/2}^T]^T$,
and the transformation functions can be given by:

$$\begin{aligned} \mathcal{F}_t(\mathbf{X}_k) = [&\mathcal{F}(\mathbf{x}_{k-T/2})^T, \mathcal{F}(\Delta\mathbf{x}_{k-T/2})^T, ..., \mathcal{F}(\mathbf{x}_k)^T, \\ &\mathcal{F}(\Delta\mathbf{x}_k)^T, ..., \mathcal{F}(\mathbf{x}_{k+T/2})^T, \mathcal{F}(\Delta\mathbf{x}_{k+T/2})^T]^T, \end{aligned} \tag{5.15}$$

where $\Delta\mathbf{x}_k$ is calculated over the contextual window length, W_D using eq. (5.8), and the TC is taken over the window length, W_C. The convergence for D-TC-INCA can be easily adapted from the convergence characteristics of INCA, and D-INCA algorithms as discussed above. Empirical convergence also observed in INCA, D-INCA, and D-TC-INCA for all the speaker-pairs. Among which the empirical convergence for one of the randomly selected speaker-pairs is shown in Figure 5.3. Cost function in all the proposed variants of INCA is different and hence, range of MSE will be different in all the cases. Still monotonically decrement in MSE sequence is clearly visible in all the cases.

5.3.3 Analysis of Phonetic Accuracies (PA)

In this work, we converted the ground truth labeling, which is at the phone-level to the frame-level labeling for the CMU-ARCTIC database [18]. The ground truth for the CMU-ARCTIC database is developed by training the speaker-dependent HMM model over 1132 utterances [18]. After alignment, using INCA algorithm and the proposed algorithm (i.e., D-INCA), if the aligned pairs are coming from the same phone-label, it is considered as hit and if not then false. From this, % Phone Accuracy (PA) is defined as [6]:

$$\% \textit{ Phone Accuracy} = \frac{\textit{Total No. Hits}}{\textit{Total No. Frames}} \times 100, \tag{5.16}$$

where *Total no. Frames* $=$ *Total no. Hits* $+$ *Total no. Falses*. Table 5.1 shows the % PA obtained using 40 non-parallel utterances from the CMU-ARCTIC database for each speaker-pairs (namely, BDL-RMS (male-male), BDL-SLT (male-female), CLB-RMS (female-male), and CLB-SLT (female-female)) using eq. (5.15). We have considered various different contextual length for calculating dynamic features in D-INCA algorithm, such as $W_D = \{3, 5, 7, 9, 11\}$.

Table 5.2: % PA analysis after alignment for different VC systems w.r.t. the different contextual window length. After [7]

Speaker-Pair	**INCA**	**D-INCA**				
		$\mathbf{W}_{D3}$	$\mathbf{W}_{D5}$	$\mathbf{W}_{D7}$	$\mathbf{W}_{D9}$	$\mathbf{W}_{D11}$
M-M	25.87	**28.94**	23.70	17.74	10.74	6.10
M-F	20.66	**24.86**	22.89	16.59	11.50	8.11
F-M	19.24	23.06	**25.37**	19.78	15.32	8.94
F-F	32.46	**36.00**	28.32	20.07	14.85	11.20

It is observed from Table 5.2 that the D-INCA (where dynamic features are calculated across three frames, i.e., W_{D3}) performs better than the INCA. Thus, dynamic features obtained over the W_{D3} is used for calculating the D-TC-INCA. We obtained clear improvement in % PA as the W_C increases in both TC-INCA and D-TC-INCA as shown in Table 5.3. In particular, we have observed on an average relative improvement of 5.39 % with our proposed D-TC-INCA over TC-INCA. In addition, it is clear from the Table 5.3 that for each considered contextual window length (namely, $W_C = \{3, 5, 7, 9, 11\}$), there is a clear improvement in % PA with

D-TC-INCA over the TC-INCA. In all the cases, the best performing system (in terms of % PA) is selected for the further development of VC system.

Table 5.3: % PA analysis after alignment for different VC systems w.r.t. the different contextual window length. After [7]

Speaker-Pair	INCA	TC-INCA					D-TC-INCA				
		$\mathbf{W}_{C3}$	$\mathbf{W}_{C5}$	$\mathbf{W}_{C7}$	$\mathbf{W}_{C9}$	$\mathbf{W}_{C11}$	$\mathbf{W}_{C3}$	$\mathbf{W}_{C5}$	$\mathbf{W}_{C7}$	$\mathbf{W}_{C9}$	$\mathbf{W}_{C11}$
M-M	25.87	28.42	28.80	30.71	31.76	**35.31**	30.90	31.85	32.51	34.33	**35.31**
M-F	20.66	23.75	26.66	29.18	28.55	**31.32**	27.09	28.32	29.33	30.98	**33.16**
F-M	19.24	24.40	26.0	27.53	27.66	**29.38**	23.51	26.16	27.49	29.56	**31.01**
F-F	32.46	36.58	39.34	40.57	43.17	**44.7**	38.2	39.92	41.26	43.99	**44.26**

5.3.4 Experimental Results

In this work, various VC systems have been developed using the aligned features obtained by INCA, and proposed dynamic INCA (D-INCA). 40 non-parallel utterances for each speaker-pairs from the CMU-ARCTIC database have been used. The state-of-the-art method, namely, Joint Density (JD) GMM-based VC has been selected among the available various VC techniques, since it uses the conditional expectation, which is the best minimum mean square error (MMSE) estimator [14, 19]. Hence, it leads to the minimum error between converted and the target spectral features. *25*-D Mel Cepstral Coefficients (MCC) (including the *0^{th}* coefficient), and *1*-D F_0 per frame (with *25* ms frame duration, and *5* ms frame shift) have been used. The number of mixture components has been varied, for example, *m=8,16, 32, 64, and 128*. The system having optimum Mel Cepstral Distortion (MCD), is selected for the subjective evaluation. Here, F_0 contour is transformed using Mean-Variance (MV) transform method [16].

5.3.5 Subjective Evaluation

We have selected Mean Opinion Score (MOS) and XAB tests for subjective evaluations of speech quality and speaker similarity (SS) of converted voice, respectively. Both the subjective tests are taken from the *19* subjects, in particular, *6* females and *13* males (with no known hearing impairments with the age between 23 to 30 years). In MOS test, subjects were asked to evaluate randomly played utterances for the speech quality (i.e., how natural is the converted voice ?) on the scale of 1 to 5 (1 very bad to 5 very good). Figure 5.4 shows the detailed MOS analysis for the developed VC systems along with 95 % confidence interval. On an average, effectiveness of the proposed D-INCA over INCA, and the proposed D-TC-INCA over the TC-INCA is visible in Figure 5.4. The poor performance of the proposed algorithm for F-F case may be due to the well known spectral resolution problem associated with female speech [12,185,186].

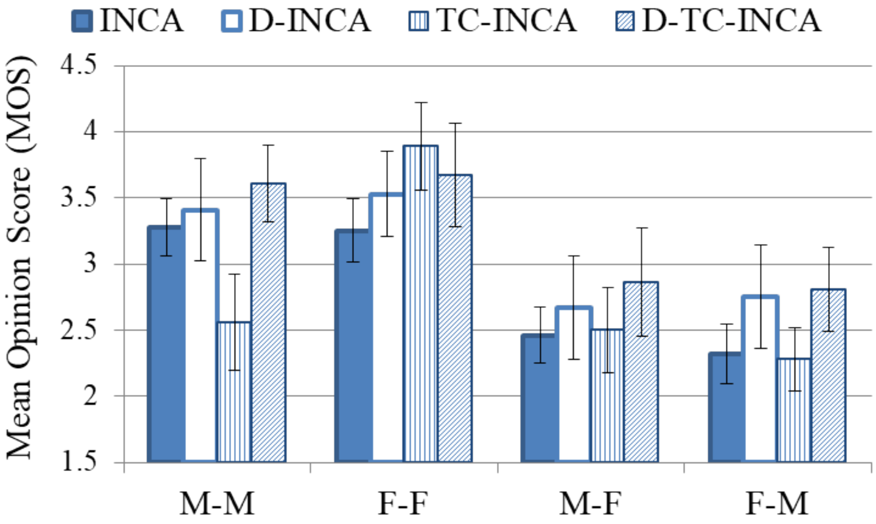


Figure 5.4: MOS analysis for VC systems. After [7].

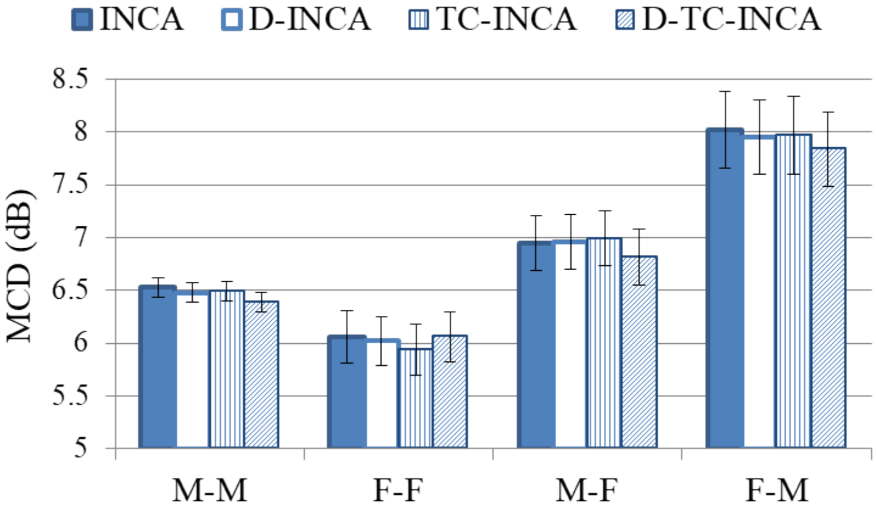


Figure 5.5: MCD analysis of VC systems. After [7].

In XAB test, the listeners were asked to select from the randomly played *A* and *B* samples (generated with INCA and TC-INCA, and the proposed D-INCA, and D-TC-INCA) based on the SS with reference to the actual target speaker's speech signal, X. We found equal preference for both the systems as subjects were unable to distinguish at all, which system is performing better in terms of SS. This result indicates an important observation that the accurate alignment may not lead to the better converted voice in terms of SS. However, it will definitely lead to the better speech quality of converted voice.

The traditional Mel Cepstral Distortion (MCD) is used for the objective evaluations of various VC systems [16]. Our proposed D-TC-INCA, and the D-INCA are performing better (i.e., having relatively lesser MCD) compared to the TC-INCA and the INCA as shown in Figure 5.5. Table 5.4 presents the analysis of Pearson Correlation Coefficient (PCC) of % PA with the MOS and the MCD. We obtained 0.36 and -0.8 correlation of % PA with the MOS and the SS, respectively.

Table 5.4: PCC between % PA and MCD with the subjective scores

PCC	**MOS**	**MCD**
% **PA**	0.36	-0.8

Next, we propose metric learning technique for finding the NN search path for alignment task in the case of non-parallel VC.

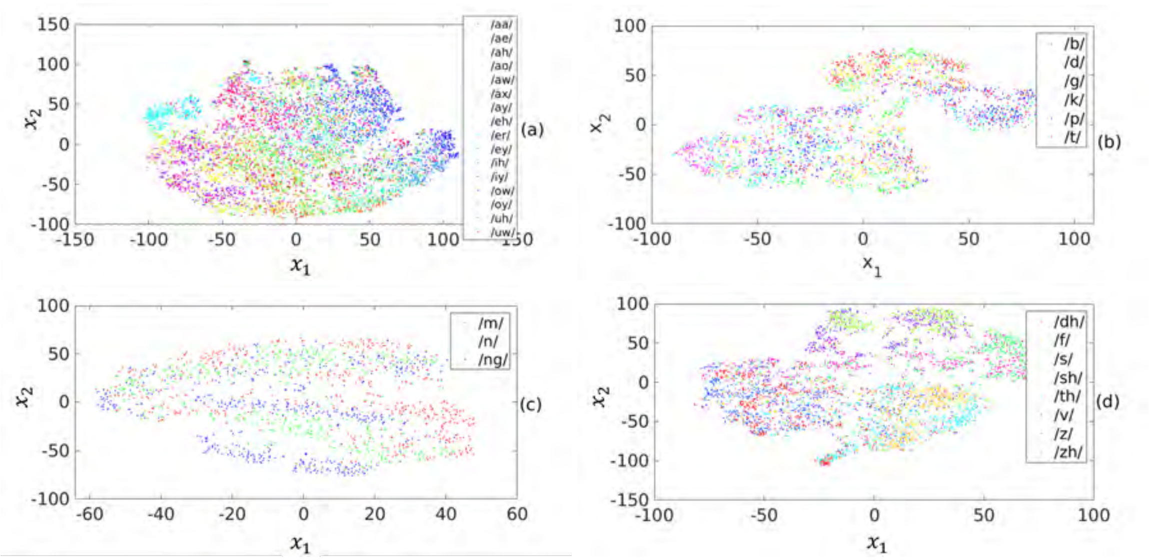


Figure 5.6: Acoustic features space visualization in 2-D using t-SNE for different speech sound classes, such as (a) vowel, (b) stop, (c) nasal, and (d) fricative. After [8].

5.4 Comparison of NN-based Alignment with CycleGAN

With the advent of the Cycle-Consistent Adversarial Network (CycleGAN), remarkable results have been reported in the case of non-parallel VC. However, in order to apply stand-alone VC techniques alignment step is unavoidable. In this work, we compare the CycleGAN architectures applied on different non-parallel alignment strategies, namely, INCA, TC-INCA, and D-TCINCA algorithm. The key objective is to investigate the genuine need of alignment step in non-parallel VC by comparing its performance with NN-based strategies.

5.4.1 Cycle-Consistent Adversarial Network for Non-parallel VC

CycleGAN [26] learns the mapping between source ($x \in X$) and target ($y \in Y$), without the need of parallel data. CycleGAN consists of two generators (G and F) and two discriminators (D_X and D_Y) (as shown by Fig. 5.7). Generator G learns the mapping function from X to Y (G: X→Y), and Generator F provides the inverse mapping function from Y to X (F: Y→X). Generator G generates distribution as $\hat{Y}$, which is passed through Generator F to give the predicted distribution of source ($\hat{X}$) and similar procedure is carried for target. Discriminator (D_X) distinguishes between real (X) and generated distribution ($\hat{X}$) of source and discriminator (D_Y) distinguishes between real (Y) and generated distribution ($\hat{Y}$) of target. Discriminator uses a back propagation method for updating weights of the generator. There are two types of cycle mapping, forward cycle consistency and backward cycle consistency, which are used to train CycleGAN. To distinguish between cycle mappings, two types of losses adversarial loss and cycle-consistency loss are introduced, in order to learn mapping of generators, G: X→Y, and F: Y→X, respectively.

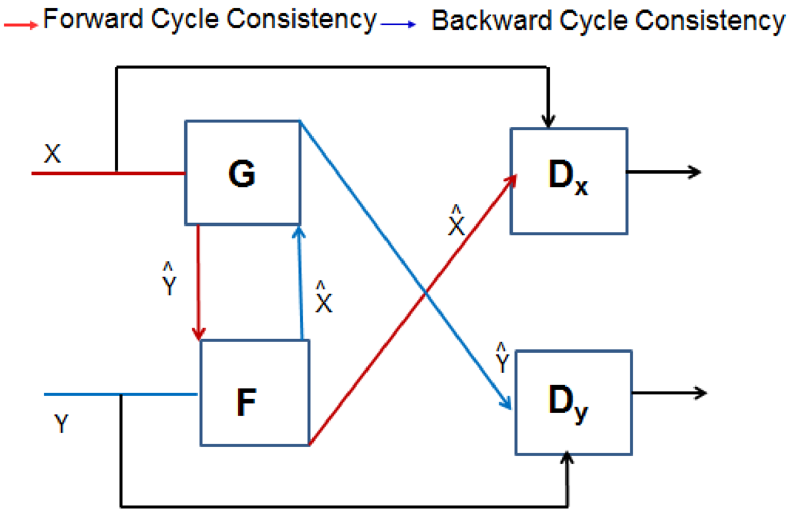


Figure 5.7: Block diagram of CycleGAN. After [26], G and F are the generators, D_x and D_y are the discriminators, X and Y are source and target speakers, respectively.

Adversarial Loss: Adversarial loss [41] measures the similarity between converted distribution (G: X→Y) from target distribution (y). Smaller this loss signifies the converted distribution is similar to the target. The objective function for Adversarial loss comprising of generator G and corresponding discriminator D$_y$ is:

$$\mathcal{L}_{adv}(G, D_Y) = \mathbb{E}_{y \sim P_{Data(y)}}[\log D_Y(y)] + \mathbb{E}_{x \sim P_{Data(x)}}[\log(1 - D_Y G(x))], \tag{5.17}$$

where $\mathbb{E}$(.) is the expectation, and P(.) is the distribution of data. Similar objective function for generator F and discriminator D$_x$, represented as ($\mathcal{L}_{adv}$(F, D$_X$)).

Cycle-consistency Loss: Cycle-consistent loss [26] function minimizes the difference between the input and output to reconstruct the input, i.e.,

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \sim P_{Data(x)}}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim P_{Data(y)}}[\|G(F(y)) - y\|_1], \tag{5.18}$$

where ||.|| means L$_1$ norm. Cycle-consistency loss preserves the contextual information of (source, target) pairs.

Complete objective function: Objective function [26] is the summation of all loss function values, i.e.,

$$\mathcal{L}(\mathcal{G}, \mathcal{F}, \mathcal{D}_\dagger, \mathcal{D}_\S) = \mathcal{L}_{adv}(F, D_x) + \mathcal{L}_{adv}(G, D_y) + \lambda \mathcal{L}_{cyc}(G, F), \tag{5.19}$$

where λ is the regularization parameter for cycle loss. During training, CycleGAN models spectro-temporal variations between different frequency bands of source and target speakers.

Optimal mapping Functions: During training, both generators tries to minimize loss in their respective objective function, while discriminator tries to maximize them. Optimal mapping function [26] of generators (G* and F*) is given by:

$$G^*, F^* = \arg\min_{G,F} \max_{D_Y, D_X} \mathcal{L}(\mathcal{G}, \mathcal{F}, \mathcal{D}_\mathcal{X}, \mathcal{D}_\mathcal{Y}, \mathcal{X}, \mathcal{Y}). \tag{5.20}$$

CycleGAN used for non-parallel VC uses additional loss function, as described:

$$\mathcal{L}_{id}(G, F) = \mathbb{E}_{y \sim P_{Data(y)}}[\|G(y) - y\|_1] + \mathbb{E}_{x \sim P_{Data(x)}}[\|F(x) - x\|_1]. \tag{5.21}$$

Identity mapping function [187] ensures that the generator finds the mapping which preserves composition between the input and output.

CycleGAN for non-parallel VC can learn from any frame pairs for training the neural network. Latent variable approach bypasses the need of speech alignment, wherein the speech signal of source is mapped to an unknown target distribution. CycleGAN uses speaker representation vectors as their latent variables, which makes the generator (G, F) capable of explaining the observation of a latent variable instead of being dependent on a pairwise transformation function [107]. Many statistical model-based VC avoids over smoothing [103]. It affects the degree of articulation of a speaker and information loss is more and is suitable only for natural speech, however not for specialized speech (whispered speech, patients suffering from neurological disorders, etc.).

5.4.2 Experimental Results

We used data from the Spoke Task of the VCC 2018 database. We developed VC systems on 16 possible speaker-pairs. *40*-dimensional (dim) Mel Cepstral Coefficients (MCCs) (including the 0^{th} coefficient), and *1*-dim F_0 were extracted from the speech signals of 25 ms window length, frameshift of 5 ms. AHOCODER is used for the analysis-synthesis of speech signals [67]. In this work, CycleGAN architecture contains the three hidden layers. Each hidden layer of generators of CycleGAN contains 512 neurons with Rectified Linear Unit (ReLU) activation, and linear activation function is applied to the output layer. The discriminators of CycleGAN also has three hidden layers, with ReLU activation function is applied to all these layers, and sigmoid activation function applied to the output layer. The model was trained for 100 epochs, using an effective batch size of 128. The parameters were optimized using ADAM optimizer, with a learning rate of 0.0001 [140].

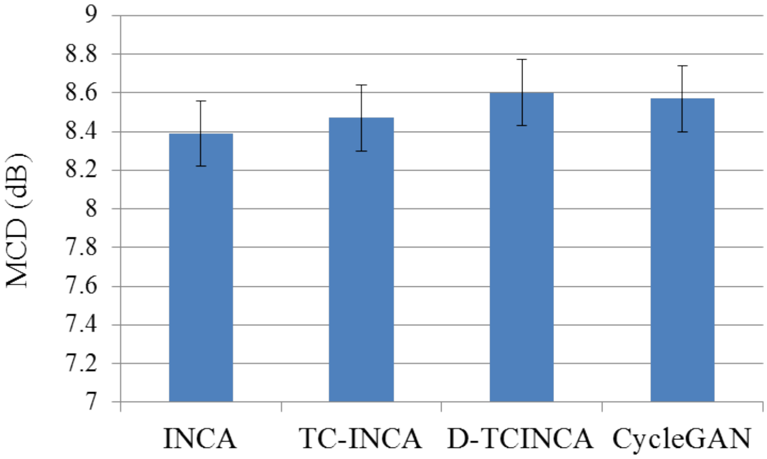


Figure 5.8: MCD analysis along with 95 % confidence interval.

Mel Cepstral Distortion (MCD) and F-ratio analysis were adopted, to carry out objective evaluation of the converted voices, which were obtained from various

VC systems. The traditional objective measure, MCD is considered here [16]. Fig. 5.8 shows MCD analysis along with the 95 % confidence interval (CI) for various VC systems. It is observed that INCA-based system shows least MCD value and CycleGAN-based system is found to be comparable with all other systems. INCA minimizes the distance iteratively by comparing source and target speakers' features using Euclidean distance. Even though MCD is considered as one of the state-of-the-art objective measures in the VC, MCD score do not correlate well with the subjective measures [5,12,14,103,107].

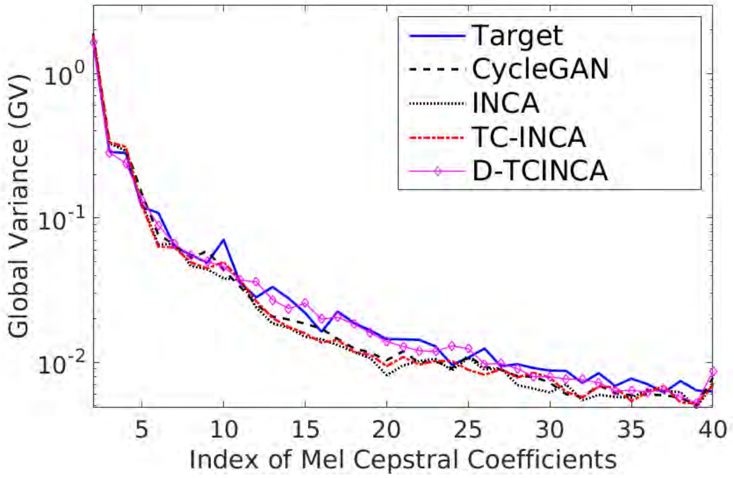


Figure 5.9: GV analysis of different VC systems.

To justify our results, we used the Global Variance (GV) analysis. It has been found that the statistical-based mapping functions using GV of the converted parameters are significantly different from the actual target speaker's parameters due to over-smoothing [16]. Fig. 5.9 shows the comparison of the GV parameters of randomly selected utterances. It is observed that the GV of the converted voice, obtained using the D-TCINCA resembles the GV of the target speakers. This indicates that the converted voices are not suffering from the *oversmoothing* issue.

Inspired from our recent study reported in [5], we propose the Fisher's F-ratio for the analysis of the converted voices in the VC context. The goal is to map the features of a source and target speakers as accurately as possible. Hence, *lower* values of F-ratio are desired between a corresponding class target speakers' voices and corresponding class for the converted voices across all the available frequency bands. Fig. 5.10 shows the F-ratio analysis for VCC 2018 database. Here, F-ratio is calculated between the class of converted voices and the actual target speakers' voices developed for all the systems using different alignment methods. Smaller the F-ratio value, less is the biasness towards the target speaker and the converted voice. It is one of the better measure in the domain of VC. It can be seen from Fig. 5.10 that there is not much difference in the F-ratio values, which makes it difficult to compare between such systems. However it is shown that the average value of INCA-based system has slightly lower F-ratio value as compared to others, across the frequency bands.

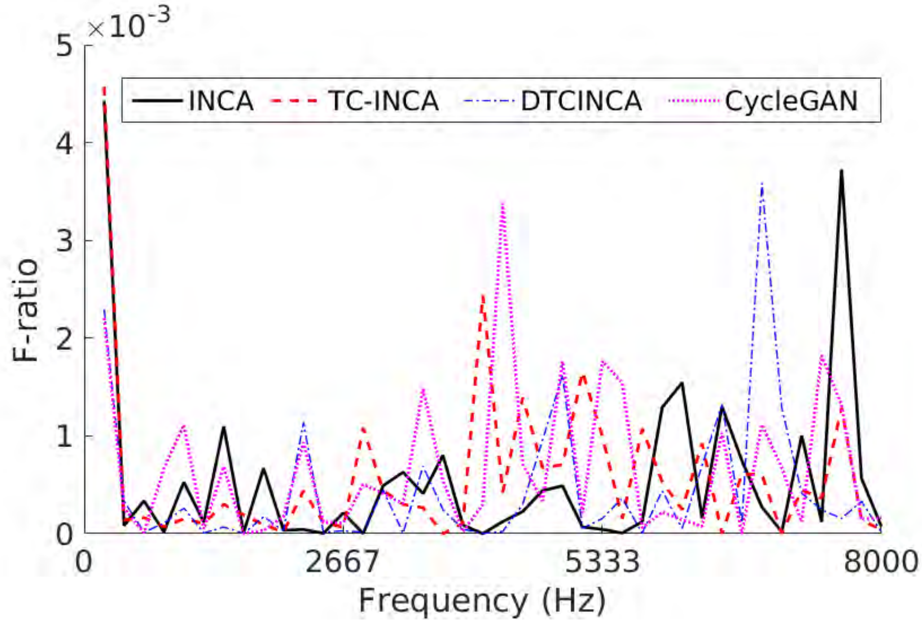


Figure 5.10: F-ratio analysis of developed VC systems.

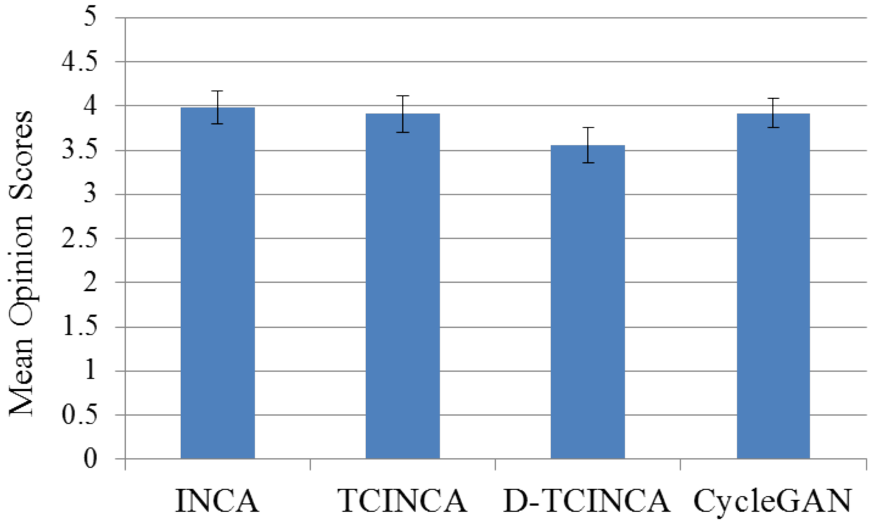


Figure 5.11: MOS analysis for speech quality along with 95 % confidence interval.

Mean Opinion Scores (MOS) is the method to evaluate the quality of naturalness in the converted voice and speaker similarity (to check similarity between synthesized speech with target speaker) is used for carrying out the subjective evaluations. High quality headphones (namely, Sennheiser HD280pro) were used for both the subjective tests. A total of 25 listeners (18 males, and 7 females in the age group 18-30 years, and no listener is known to have hearing impairments) have participated in both the subjective tests. Four utterances of each VC system were played and each subjects were asked to rate the speech quality and speaker similarity on 5-point scale. MOS scores obtained for both the speech quality and speaker similarity are represented in Fig. 5.11 and Fig. 5.12, respectively. We can observe that the CycleGAN-based system is performing relatively well as compared with different alignment (non-parallel) methods.

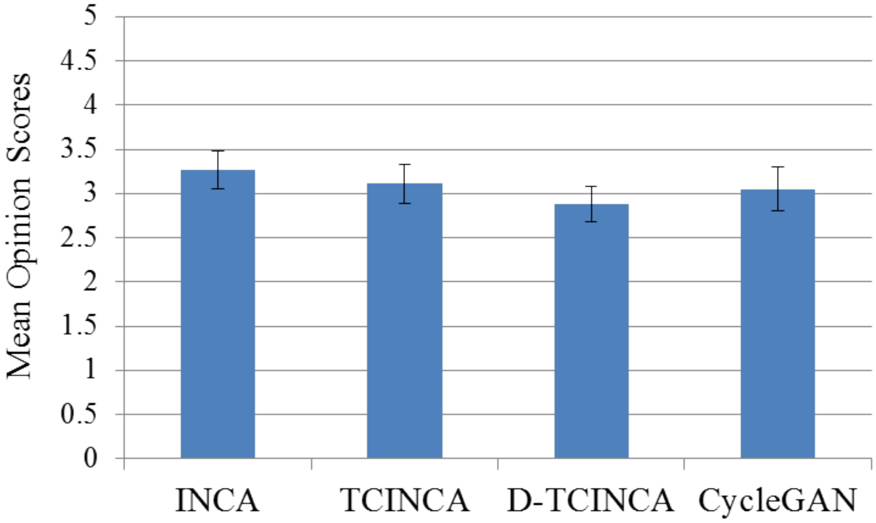


Figure 5.12: MOS analysis for speaker similarity along with 95% confidence interval.

5.5 Metric Learning for Alignment Task

5.5.1 Motivation for Metric Learning in VC

In the literature, lower Phonetic Accuracy (PA) is reported after the alignment step [78]. To further investigate the possible reasons for this, we apply *t*-stochastic neighbor embedding (t-SNE) visualization technique to the acoustic space of the source and the target speakers [188]. We have taken one of the available speaker-pairs (namely, BDL-RMS (M-M)) from the CMU-ARCTIC database [18]. The acoustic space for a vowel, stop, nasal, and fricative is shown in Figure 5.6. We can clearly see that the same phoneme uttered by the two speakers does not lie in the neighborhood in Euclidean space, rather they are spread across the 2-D acoustic space. This is primarily due to the difference in vocal tract system (i.e., size and shape), and excitation source (difference in size of the glottis, vocal fold mass, tension in the vocal folds and hence, the manner in which glottis opens or closes, i.e., the glottal activity) across the speakers (Chapter 3, pp. 59) [17]. This motivated us to define a new metric that represents the acoustic feature space. Hence, we propose to use the learned metric for finding the NN pairs in the second step of INCA [6].

5.5.2 Metric Learning

The metric learning is concerned with the learning of a distance function w.r.t. a particular task. Metric learning has been shown to be extremely useful when used along with the NN methods [189]. The metric learning techniques can be broadly classified into linear (which uses Mahalanobis distance) *vs.* nonlinear approaches [189]. Let $X = [x_1, x_2, ..., x_n]$ be the matrix of all the data points. The mapping $d : X \times X \rightarrow \mathbb{R}$ is called a *metric* if it satisfies following four conditions [178]:

1. $d(x_i, x_j) \geq 0$ (non-negativity),
2. $d(x_i, x_j) = 0 \Leftrightarrow x_i = x_j$ (identity of indiscernible),
3. $d(x_i, x_j) = d(x_j, x_i)$ (symmetry),
4. $d(x_i, x_j) \leq d(x_i, x_r) + d(x_r, x_j)$, where $\forall x_i, x_, x_r \in X$ (triangle inequality).

If condition (2) is dropped then the mapping is called as *pseudo* metric [178]. In particular, distance metric is defined through inner product space. For example, $d^2(x,y) = ||\langle x-y, x-y \rangle|| = x^T y$. Hence, in general a distance metric is defined as:

$$d_A(x,y) = (x-y)^T A(x-y). \tag{5.22}$$

If $A = \Sigma^{-1}$ then distance is called as the Mahalanobis distance [155]. Here, Σ is covariance matrix of the data. In most of the cases, true covariance is unknown and hence, the sample covariance is used. Here, A must be positive-semidefinite (PSD) (i.e., $A \succeq 0$, where $\succeq$ notation is used to indicate positive semidefinite) to satisfy the metric definition. Furthermore, if A is PSD then it can be factorized as $A = G^T G$ that leads to $d_A(x,y) = ||Gx - Gy||_2^2$ (where $|| \cdot ||$ is the L^2 norm). Hence, the idea behind learning the metric can be considered as the learning of global linear transformation. The idea behind learning Mahalanobis metric was proposed by Xing *et. al.* in [190]. Here, the desired metric should give minimum squared distance for the pairs $(x_i, x_j) \in \mathcal{S}$ (where S is a set of similar pairs) with constraint, $\sum_{(x_i,x_j) \in D} d_A(x_i, x_j) \geq 1$, where D is set of dissimilar pairs. Here, the objective function is given by [190]:

$$\arg\min_{A} \sum_{(x_i,x_j)\in\mathcal{S}} ||x_i - x_j||_A^2, \tag{5.23}$$

$$\text{subject to the condition} \sum_{(x_i,x_j)\in\mathcal{D}} ||x_i - x_j||_A^2 \geq 1, \quad A \succeq 0. \tag{5.24}$$

The above mentioned objective function is linear and both the constraints are convex and hence, the convex optimization algorithm can be applied to get global optimal solution. In particular, the gradient descent and the idea of iterative projections can be used to solve the above mentioned convex optimization problem [190]. Weinberger *et. al.* proposed the LMNN technique that uses the relative distance constraints, which is one of the most popular and state-of-the-art metric learning technique in the literature [27, 191]. The main aim of LMNN technique is that the given feature should have the same label as its neighbors, while the features that are having different labels should be distant apart from the given feature. The key idea behind the LMNN is illustrated in Figure 5.13.

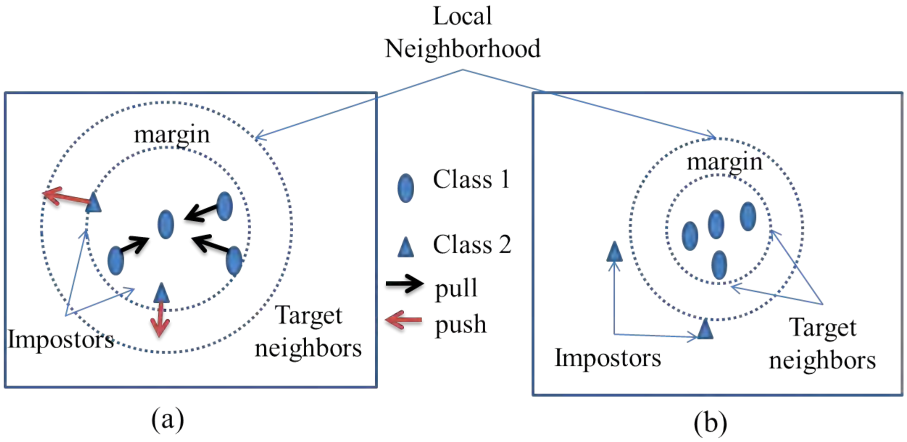
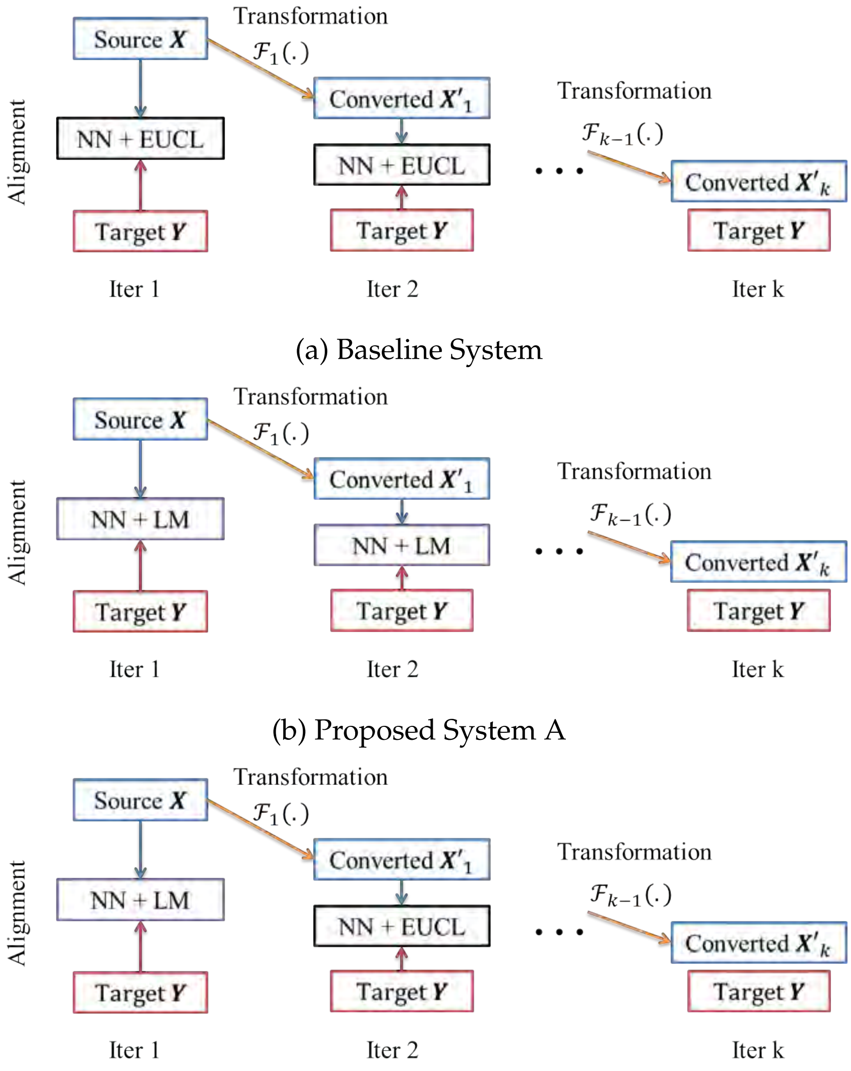


Figure 5.13: Schematic representation of LMNN technique (a) before, and (b) after applying the LMNN technique. Adapted from [27].

Here, the target neighbors refer to the features that have similar label, and impostor is also the neighbor feature vector. However, it is having different label. The goal of LMNN technique is to minimize the number of impostors via relative distance constraint. The objective function is given by [27]:

$$\arg\min_{A\succeq 0} \sum_{(i,j)\in\mathcal{S}} d_A(x_i, x_j) + \lambda \sum_{(i,j,k)\in\mathcal{R}} [1 + d_A(x_i, x_j) - d_A(x_i, x_k)], \tag{5.25}$$

where $\mathcal{R}$ is the set of all triplets (i, j, k) such that x_i and x_j are the target neighbors, and x_k is the impostor.



(a) Baseline System

(b) Proposed System A

(c) Proposed System C

Figure 5.14: Schematic representation of (a) baseline, (b) proposed system A, and (c) proposed system C. Proposed system B is not shown here, since it applies the baseline technique to the transformed features obtained via the LM, and hence, similar to (a). EUCL: Euclidean metric, LM: Learned metric. After [8].

In this work, we have used TIMIT database for estimating the learned metric as the manual phone-annotations are available, which is obtained from the highly trained human annotators [192]. We used the full phone-set label. We randomly selected a small subset of the database to learn the metric using LMNN technique. We extracted *25-D* Mel Cepstral Coefficients (MCC) per frame (with *25* ms frame duration, and *5 ms* frameshift). In this thesis, we globally learn the metric for the spectral features, and use this learned metric for calculating NN feature-pairs. We considered three possible approaches to use the learned metric in the INCA. Schematic representations of the various approaches are given in Figure 5.14. Proposed system A uses the learned metric in each iteration of the baseline INCA as shown in Figure 5.14 (a) and (b). On the other hand, the proposed system C uses the learned metric only at the iteration I in the INCA as shown in Figure 5.14 (c). Proposed system B first applies the global transformation that is learned via metric learning to the spectral features obtained from both the speakers, and then the baseline INCA is applied to the transformed features.

5.5.3 Experimental Results

We have used CMU-ARCTIC database due to the availability of the phone annotations that is obtained using speaker-dependent hidden Markov model (HMM) trained over 1132 utterances [18]. Here, we converted the reference phone annotations to the frame-level labeling. In this thesis, 40 non-parallel utterances from each speaker-pair have been used to develop VC system using the aligned features obtained via the baseline and proposed techniques. The state-of-the-art methods, namely, Joint Density Gaussian Mixture Model (JDGMM)-based VC has been selected among the available various VC techniques [19]. The JDGMM-based method is selected since it uses conditional expectation, which is the best minimum mean square error (MMSE) estimator [193]. Hence, it leads to the minimum error between converted and the target spectral features. *25*-D MCC and *1*-D F_0 per frame (with *25* ms frame duration, and *5* ms frameshift) have been extracted using AHOCODER. The number of mixture components has been varied, for example, *m=8,16, 32, 64, and 128*. The system having optimum Mel Cepstral Distortion (MCD), is selected for the subjective evaluation. Here, Mean-Variance (MV) transformation has been used for F_0 conversion [16].

5.5.4 Analysis of Phonetic Accuracies

Figure 5.15 shows the accuracy of alignment obtained using three proposed techniques. It is observed that there is on an average 0.71 % relative reduction, and 0.07 % relative improvement (in the PA) w.r.t. the baseline using the proposed method A and B, respectively. It is possibly due to fact that this metric is globally learned for the entire TIMIT database. The broad phonetic classes, such as vowel, stops, fricatives, nasal, etc. will behave very much differently in acoustic space due to the different manner of articulation required to produce these speech sounds [17]. Since the metric is learned for the true acoustic features and not for the intermediate converted acoustic features, we propose technique C, which is found to be performing consistently better (with on an average 7.93 % relative improvement in PA) than the INCA.

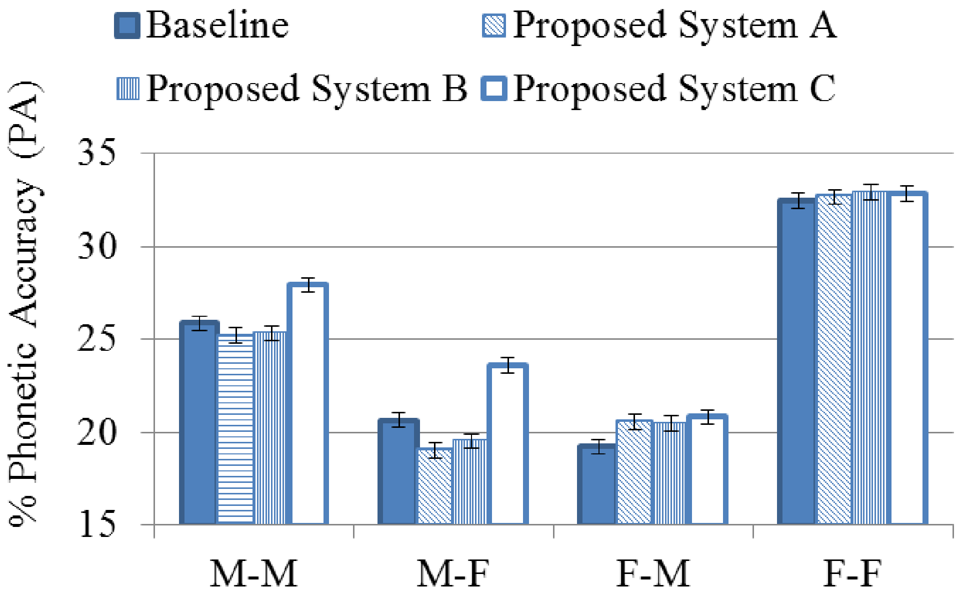


Figure 5.15: PA of different initialization techniques for non-parallel VC systems. After [8].

5.5.5 Subjective and Objective Evaluations

Mean Opinion Score (MOS) and ABX tests have been performed for measuring speech quality, and speaker similarity (SS) of the converted voices, respectively. Both the subjective tests are taken from the *16* subjects (*5* females and *11* males with no known hearing impairments and with the age variations between 18 to 29 years) from the total of *288* samples.

Table 5.5: MOS analysis for the naturalness of converted voices. Number in the bracket indicates a margin of error corresponding to the 95 % confidence intervals for VC systems. After [8]

	M-M	**M-F**	**F-M**	**F-F**
Baseline	3.06 (0.27)	2.41 (0.29)	**2.66** **(0.28)**	3.5 (0.26)
Proposed System C	**3.31** **(0.29)**	**2.81** **(0.22)**	2.53 (0.21)	**3.5** **(0.25)**

The result of *5*-point MOS test are shown in Table 5.5 along with the 95 % confidence intervals. It can be seen from Table 5.5 that the proposed system C is more preferred than the baseline in terms of speech quality (i.e., naturalness) for the VC (except in the case of F-M). In ABX test for SS, the listeners were asked to select from the randomly played *A*, and *B* samples (generated with the baseline, and the proposed system C) based on the SS with reference to the actual target speaker's speech signal X. Eight samples for ABX test were taken from both the approaches. All the subjects have given *equal* preference to both the systems. This result indicates that the accurate alignment may not lead to the better converted vSTMoice in terms of SS. However, it may lead to the better speech quality of converted voice.

Table 5.6: MCD analysis. Number in bracket indicates the margin of error corresponding to the 95 % confidence intervals. After [8]

	M-M	**M-F**	**F-M**	**F-F**
Baseline	6.53 (0.34)	6.95 (1)	8.02 (1.29)	6.06 (0.93)
Proposed System C	**6.41** **(0.09)**	**6.76** **(0.26)**	**7.85** **(0.34)**	**6.02** **(0.24)**

Table 5.7: PCC of PA and MCD with the subjective scores, MOS and SS. After [8]

PCC	**MOS**	**SS**
PA	0.96	0.37
MCD	-0.3	0.10

The traditional Mel Cepstral Distortion (MCD) is used here as an objective measure [16]. It can be seen from the Table 5.6 that our proposed system C is performing better than the baseline system in all the cases. Table 5.7 presents the analysis of Pearson Correlation Coefficient (PCC) of PA and MCD with the MOS and the SS. It is clear from the Table 5.7 that the PCC between PA and MOS is 0.96, i.e., PA is having more correlation with the MOS than with the MCD. This clearly indicates that better alignment will lead to better speech quality. On the other hand, PCC between PA and SS is less compared to the PCC between PA and MOS. For the case of MCD, PCC ideally should be -1 since lesser value of MCD means that the system is performing better than the given VC systems (that are having higher values of MCD). It is clearly seen that the traditional MCD is *not* correlating well with the MOS and the SS. Less correlation between MCD and the subjective scores have also been reported in the VC literature (which may be due to the inability of MCD to capture the characteristic of sophisticated human perception of hearing, as discussed in sub-section 4.4.3 from the Chapter 4) [12]. Next, in order to exploit phonetic information for the NN-based alignment technique, the Spectral Transition Measure (STM) technique is proposed to estimate phonetic boundaries in the context of non-parallel VC.

5.6 Phone Aware Nearest Neighbor Technique using Spectral Transition Measure (STM)

Lower % Phonetic Accuracy (PA) has been reported in the VC literature after the NN-based alignment techniques [6, 14, 78]. Hence, if the phone boundaries corresponding to the non-parallel data are available, then the NN can be applied among the features corresponding to the same phones. Estimating the phone boundaries is more challenging due to the coarticulation phenomenon of speech sounds, which leads to splitting or disappearing of current sound due to merging with or interference of the adjacent sounds (primarily due to local *vs.* global coarticulation) [40]. If the text corresponding to the training data is available, Hidden Markov Model (HMM) can be used to identify the correct phone boundaries or recently proposed speaker-independent phone posterior probability features can also be used [82, 104]. However, these methods require a large amount of training data to train HMM or develop Automatic Speech Recognition (ASR), which is difficult due to the unavailability of a large amount of transcribed training utterances from the target speaker in most of the real-world VC applications.

In this work, we propose to exploit computationally simple Spectral Transition Measure (STM)-based alignment technique that does not require any apriori training data for estimating the phone boundaries. Earlier, the STM-based alignment techniques had also been used for identifying the syllable and phone boundaries for the low-resource languages [28, 194–198]. In this work, we propose a simple and practical way of utilizing the phone information via STM algorithm for the alignment task in non-parallel VC. The phone boundaries estimated using the STM algorithm are then applied to the NN method (i.e., phone-aware NN) to find the alignment between the source and target speakers' spectral features.

5.6.1 Proposed STM-based Alignment

Figure 5.16 shows an example of manual phone boundaries obtained for a CMU-ARCTIC speech utterance from two speakers. Even if both the speakers have spoken the utterances with different speaking rate, still the phone boundaries occurs at the spectral transition locations as shown in Figure 5.16.

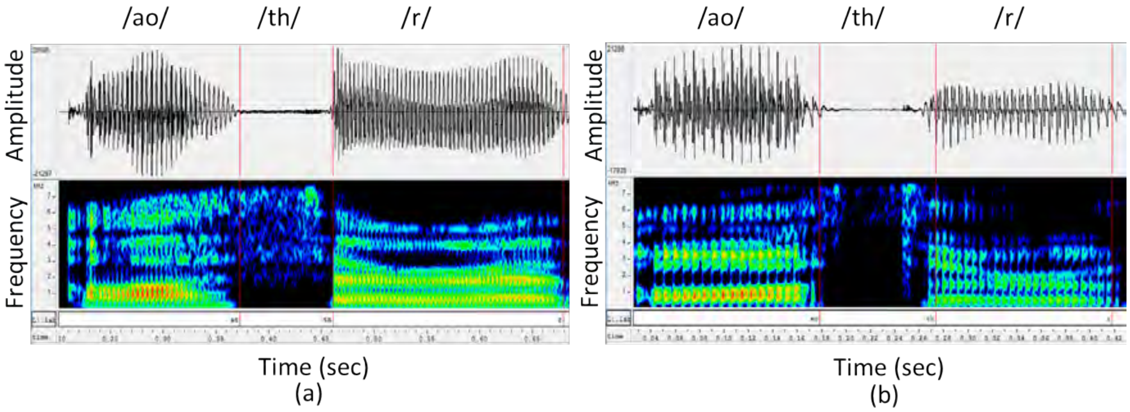


Figure 5.16: Speech signal, and its corresponding spectrogram for English utterance, namely, “author" from CMU-ARCTIC database from (a) male, and (b) female speaker. After [28].

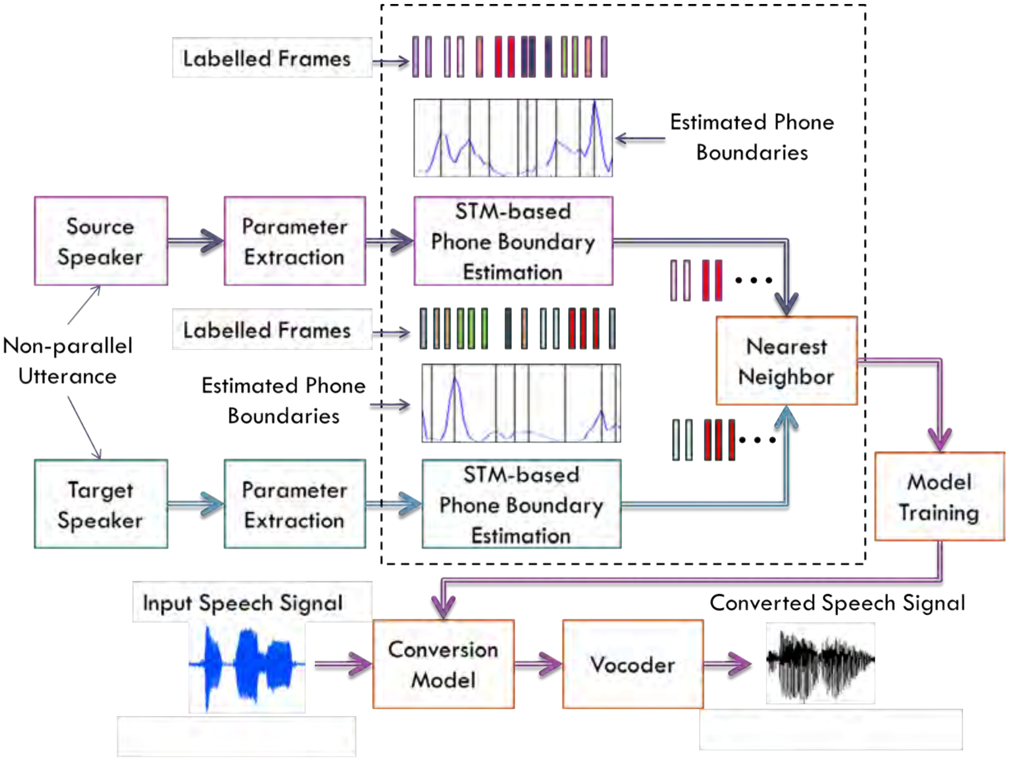


Figure 5.17: Block diagram of the proposed STM-based non-parallel VC system. The contribution of this thesis is indicated via dotted box. After [9].

It is often observed that two human annotators can never identify exactly the same phone boundaries. In addition, obtaining the manual phonetic segmentation on the given speech corpus is extremely tedious and time-consuming. Furthermore, it requires highly trained human annotators, which makes this process very costly. In real-time VC, it is impossible to do manual segmentation as soon as one gets the training data from the target speaker. Hence, there is a need for automatic segmentation algorithm for the alignment task of VC. Speech signal consists of various basic speech sound units called as *phonemes*. These sounds and their features differ both in time, and spectral characteristics. The dynamic features, such as, temporal envelope characteristics and the change of distribution of spectral energy will play a key role to distinguish major phonetic categories, such as vowels, nasals, stops, fricatives, etc. [53, 199, 200]. These rapid changes in amplitude, and spectrum are apparently represented in the discharge pattern of auditory-nerve fibers (ANFs) [201]. This spatio-temporal pattern of auditory-nerve activity has shown to contain pointers to the regions of rapid changes that captures the important phonetic (transitional) information [201, 202]. In addition, the earlier studies reported that the neurons present in the auditory cortex poorly respond to the steady-state stimuli, whereas they have high auditory sensitivity for the *transitional* sounds [203]. Thus, STM exploits spectral variations to detect the phone boundaries. Figure 5.17 shows the block diagram of proposed approach.

Here, Mel Frequency Cepstral Coefficients (MFCC) have been used for capturing these spectral transitions across the consecutive phones. First, the speech signal is pre-processed using the silence removal technique. In order to estimate the phone boundaries, we have followed the same experimental setup as suggested in [28, 204]. *10*-dimensional (D) MFCC feature vectors (including 0^{th} coefficient) are first extracted (with 30 ms window and 10 ms frame rate) as suggested in [28, 198, 204]. After computing the MFCC, STM is used to capture the spectral variations between the two consecutive phones. The STM, at the i^{th} frame, can be computed as [180]:

$$C(i) = \frac{\sum\limits_{l=1}^{L} a_l^2(i)}{D}, \tag{5.26}$$

where $C(i)$ is the STM calculated at a given frame i, D is the dimension of the spectral features vectors, a_l's are the regression coefficients, which are the rate of change of spectral features. The a_l is given by [180]:

$$a_l(i) = \frac{\sum\limits_{j=-I}^{I} MFCC_l(j+i) \cdot j}{\sum\limits_{j=-I}^{I} j^2}, \tag{5.27}$$

where j is the frame index, and I indicates the number of frames (on both side of the current frame) used to compute the linear regression coefficients. Finally, peak detection algorithm is used to estimate the phone boundaries. A number of estimated boundaries may not be equal to the number of phones in a given utterance and hence, the proposed algorithm is used to get the exact number of estimated boundaries equal to the number of phones. Here, we have taken $I = 2$ for 10 ms frame shift. Hence, the value of C in eq. (5.21) is computed over an interval of 40 ms. A larger interval results in the missing of some phone boundaries, whereas the shorter interval results in a false estimate of the phone boundaries. Previous studies presented the effectiveness of the algorithm in the context of a tolerance interval [28,195,198]. It means that if the detected phone boundary is within the tolerance interval (namely, 5 ms, 10 ms or 20 ms) of the ground truth, then it is considered as a true estimated boundary else detected boundary is considered as false.

In the context of VC task, we consider 0 ms tolerance interval, i.e., the frames are aligned based on these estimated phone boundaries. The exact locations of the phone boundaries will determine the corresponding pairs among which NN technique will be applied. Hence, we converted the phone-level labeling to the frame-level labeling, and compared it with the corresponding frame-level labeling of the ground truth. If both the estimated and the ground truth frame-level label are found to be the same, it is considered as hit and if not then false. From this, % Phonetic Accuracy (PA) is defined as [72]:

$$\%\ Phonetic\ Accuracy = \frac{Total\ No.\ Hits}{Total\ No.\ Frames} \times 100, \tag{5.28}$$

where *Total No. of Frames = Total No. of Hits + Total No. of Falses*. Table 5.8 shows the average % Phonetic Accuracy (PA) for TIMIT and CMU-ARCTIC database (BDL, CLB, RMS and SLT speakers) using eq. (5.23). The ground truth for the TIMIT database is developed by highly trained annotators [192]. On the other hand, reference phone annotations for the CMU-ARCTIC database are obtained via training of speaker-dependent HMM model over 1132 utterances [18].

Table 5.8: % Phonetic Accuracy (PA) of STM algorithm. After [9]

	TIMIT Database	**CMU-ARCTIC Database**
% **PA**	27.53	31.63

5.6.2 STM with Nearest Neighbor (NN)

The proposed STM algorithm is shown in Algorithm 2. Once the phone boundaries are estimated for source and the target speakers' training data, the nearest neighbor (NN) technique is applied to find correspondence between spectral features of the source and the target speakers among the same labels of frames estimated using the STM algorithm. For the baseline method, we selected the INCA [6,205].

Algorithm 2 Proposed STM-based Algorithm for VC task. After [9]

```
Input: Speech wav file and the corresponding utterance from both the source and target speakers.
Preprocessing of silence removal from the speech file.
MFCC feature extraction from the speech file.
Extraction of a STM contour at the frame-level.
NE: Number of estimated STM boundaries using peak detection.
NG: Number of ground truth boundaries from a given utterance.
NI: Number of insertions.
ND: Number of deletions. $NE < NG$
$NI = NG - NE$. $NI \neq 0$
find two furthest neighbors estimated boundaries.
Insert boundary based on average phone duration.
$NI = NI - 1$
done $NE > NG$
$ND = NE - NG$. $NI \neq 0$
Find the two nearest neighbors estimated boundaries.
Merge them by selecting either left or right.
ND=ND-1
done
end
Estimate boundaries for both source and the target speakers' training speech utterances.
Apply Nearest Neighbor (NN) for a given phoneme.
Estimate the unique aligned pairs between source and the target speakers from the NN path.
Train the mapping function using the obtained aligned pairs.
```

Figure 5.18 shows the % Phonetic Accuracy (PA) calculated using the proposed technique, and the baseline algorithm for four different CMU-ARCTIC database pairs, such as BDL-RMS (male-male), CLB-SLT (female-female), BDL-SLT (male-female), and CLB-RMS (female-male). From Figure 5.18, it is clear that the proposed STM+NN technique is giving better performance compared to the baseline algorithm in all the four cases. The proposed STM+NN algorithm is having on an average 13.67 % relative improvement in % PA compared to the baseline algorithm. It has been shown that the conversion of vowel affects the quality of converted voices in the VC literature [14].

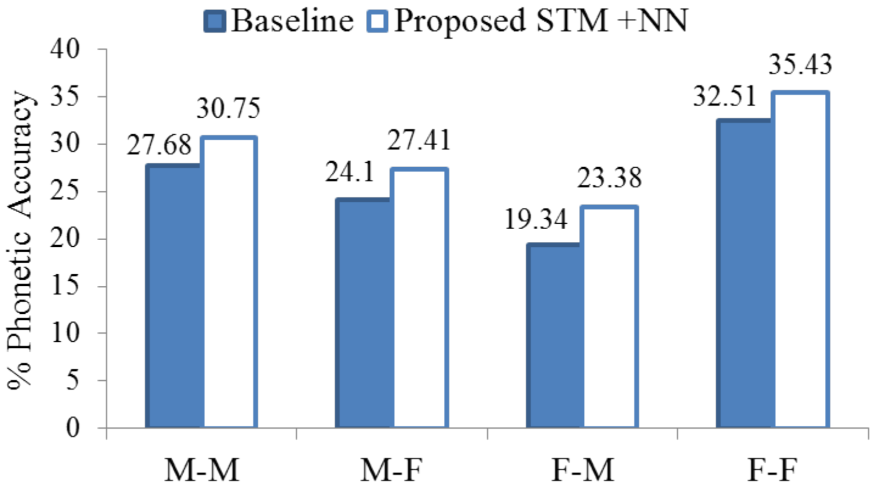


Figure 5.18: % Phonetic Accuracy (PA) of alignment techniques for different non-parallel VC systems. After [9].

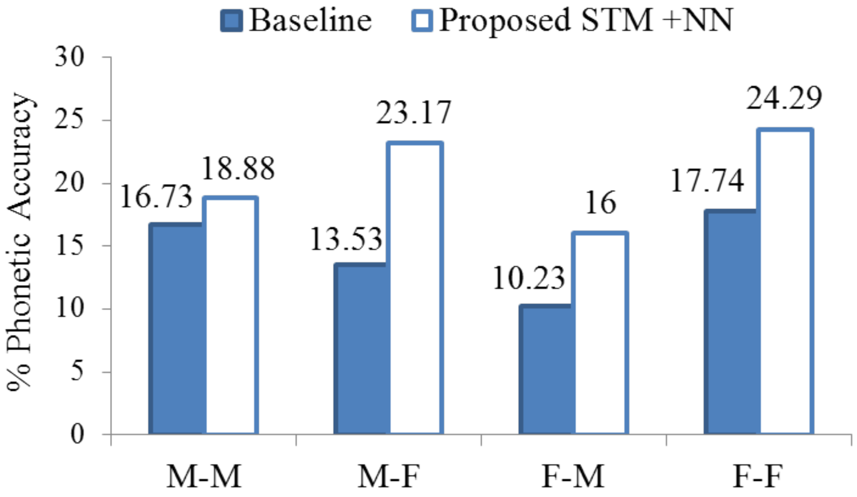


Figure 5.19: % Phonetic Accuracy (PA) for vowels using different alignment techniques for different non-parallel VC systems. After [9].

Figure 5.19 shows the % PA for only vowel sounds using the baseline and proposed alignment method. The effectiveness of the STM to detect the vowel sounds are indeed observed from the Figure 5.19. In particular, we obtained on an average 44.36 % absolute increment in correct vowel detection. In addition, the baseline algorithm takes more time compared to the proposed alignment technique due to the iterative computation (approximately, number of iterations times the time taken by the proposed STM+NN algorithm [9]).

5.6.3 Experimental Results

In this work, various VC systems have been developed to measure the effectiveness of the proposed alignment task. We have used 40 non-parallel utterances for each speaker-pairs from the CMU-ARCTIC database. Among the available VC techniques, the state-of-the-art methods, namely, Joint Density (JD) GMM-based VC [19], and BiLinear Frequency Warping plus Amplitude Scaling (BLFW+AS) [42] have been selected. The JDGMM-based method is selected, since it uses conditional expectation, which is the best minimum mean square error (MMSE) estimator. Hence, it leads to the minimum error between converted and the target spectral features. In addition, BLFW+AS has been selected due to its available parametric formulation [42]. *25*-D Mel Cepstral Coefficients (MCC), and *1*-D F_0 per frame (with *25* ms frame duration, and *5* ms frame shift) have been extracted. The number of mixture components have been optimized from the set *m=8,16, 32, 64, 128*. System having an optimum MCD, is selected for subjective evaluation. Here, fundamental frequency (i.e., F_0) contour is transformed using Mean-Variance (MV) transform method [16]. The AHOCODER is used for the analysis-synthesis [67].

Table 5.9: MOS analysis. Here, we obtained 0.175 margin of error corresponding to the 95 % confidence interval. After [9]

	Baseline	**Proposed STM + NN**
MOS	3.08	**3.51**

Table 5.10: XAB test analysis for speaker similarity. After [9]

	Baseline	**Proposed STM + NN**	**Equal Preference**
Preference Score (%)	21.25	**24.7**	54.05

The Mean Opinion Score (MOS), and XAB tests have been selected to evaluate speech quality (i.e., naturalness), and Speaker Similarity (SS) of the converted voice, respectively. Total *17* subjects (*4* females and *13* males without any hearing impairments with the age between 17 to 28 years) took part in both the tests. Subjects were asked to evaluate the randomly played utterances from both the GMM and BLFW-based VC techniques for the speech quality on the scale of 1 (very bad) to 5 (very good).

It can be seen from Table 5.9 that proposed STM+NN is more preferred than the baseline system in terms of *speech quality*. The result clearly indicates that the accurate estimation of phone boundaries is indeed helping the NN-based technique to obtain high quality quality converted voice. In XAB test, subjects were asked to select from the randomly played *A* and *B* samples (generated with the baseline, and the proposed STM+NN algorithm) which one is more similar with the target speaker in terms of *speaker similarity* (SS) with reference to the actual target sample X. In addition, the subjects can select equal preference for the cases, where the samples are perceptually similar in terms of speaker similarity. Samples for XAB test, were taken from both the GMM and BLFW-based systems. It is observed from the Table 5.10 that the proposed STM-based alignment technique is preferred 3.45 (= 24.7-21.25)% times more for the speaker similarity of the converted voice. However, 54.05 % times subjects have given equal preference to both the systems.

5.7 Chapter Summary

In this chapter, we presented different NN-based strategies related to the alignment task in the context of non-parallel VC. In particular, INCA algorithm is discussed along with the research issues, and the proposed approaches. Furthermore, formal convergence theorem has been proposed for the INCA algorithm. In particular, it has been shown that the INCA algorithm will converge monotonically to a local minimum in mean square error (MSE) sense. Furthermore, we also present the reason of convergence in MSE sense (opposed to pointwise) in the context of VC task by considering an example from twins, who look and sound perceptually very similar. Then, the idea of using dynamic features along with static features to calculate the NN-aligned pairs in both the INCA and TC-INCA algorithms is proposed (since the dynamic features are known to play a key role to differentiate major phonetic categories).

One of the key issues in the INCA is that it tries to minimize the Euclidean metric among acoustic features for the time alignment. However, the same phoneme uttered by the two speakers may not have the minimum Euclidean distance. Hence, we proposed to exploit metric learning technique for finding NN in the INCA than state-of-the-art Euclidean distance. Furthermore, we also proposed to use our learned metric only for the initial iteration of INCA (since the metric is learned for the actual acoustic features). Therefore, during other iterations in the INCA, intermediate converted features may not represent the true acoustic features.

Next, in order to exploit the phonetic information in NN-based alignment techniques, the novel STM+NN-based algorithm has been proposed for the task of alignment in the case of text-independent VC. The key advantage of the proposed method is that it does not require any a priori training data to estimate the phone boundaries. The % PA obtained after alignment using STM algorithm is found to be better compared to the baseline NN-based alignment technique in all the cases. In particular, % PA obtained for the vowel speech sound class is more using the proposed STM algorithm. The better performance in the alignment task has resulted positively in the context of subjective test for the developed VC systems using proposed approach. In the next chapter, novel VC architecture will be discussed, which removes the need of alignment step and hence, it can be applied to parallel as well as non-parallel VC tasks.

CHAPTER 6

Strategies to Overcome Alignment Step

6.1 Introduction

Alignment strategies and its impact on the quality of the converted voice has been discussed for parallel and non-parallel VC in Chapter 4 and Chapter 5, respectively. The alignment techniques will generate one-to-many and many-to-one feature pairs in parallel as well as non-parallel VC task [6, 39]. In addition, if a word spoken by a source speaker is repeated by the target speaker with different variations, it will generate such pairs. Furthermore, if the same word is repeated several times, this repetion will result in the different speech pattern. This also generates such kind of pairs. Directly learning the relationship in the presence of such pairs is very challenging. These one-to-many and many-to-one pairs will affect the learning of the mapping function and thus, results in the muffling and oversmoothing effect in VC [206].

The earlier approaches used context-dependent information to overcome this issue [206]. Recently, equalizing formant locations using Dynamic Frequency Warping (DFW) was proposed to tackle these issues [207]. In addition, some of the approaches proposed to filter out such pairs from the training [49, 208]. However, loosing number of pairs will not be useful in the case, where the amount of training data is small. There is also an attempt in the past to use pre-stored speakers parallel data to train initial model in the case of non-parallel VC task [209]. Furthermore, adaptation [81, 210] and the model-based [101, 107–109] approaches have been proposed to avoid alignment step, which also help to solve one-to-many pairs-related issues. Recently, Phonetic Posteriorgram (PPG) (which are believed to be speaker-independent representations) have been proposed that consider two-stage mapping [82, 104]. However, it requires a huge amount of labeled speech data to train the Automatic Speech Recognition (ASR) systems for estimating PPG. Since the training data is small in most of the applications of non-parallel VC. Hence, we propose to exploit unsupervised technique for computing speaker-independent posterior features.

In this chapter, we propose to avoid alignment step by using the two separate Deep Neural Networks (DNNs), where one DNN will map source speaker's spectral features to the speaker-independent feature representations, and the another DNN will map these speaker-independent feature representations to the particular target speaker's spectral features. Here, we propose to use unsupervised Vocal Tract Length Normalization (VTLN)-based warped Gaussian Posteriorgram (GP) and Inter-Mixture Weighted (IMW) GP features as the speaker-independent representations.

6.2 Unsupervised Speaker-Independent Posterior Representations

Earlier Gaussian Mixture Model (GMM) posteriorgram, and PPG were used as features for phoneme classification, and template matching-based ASR [211–214]. In this work, three different types of speaker-independent posterior features have been considered, namely, phonetic posteriorgram, GMM posteriorgram, and Vocal Tract Length (VTL)-warped posteriorgram.

6.2.1 Phone Posteriorgram (PPG)

PPG contains the posterior probability for each phonetic class obtained for a given speech signal [82]. Most of the non-parallel VC systems are having less number of utterances for training and hence, developing the ASR system for obtaining PPG is very difficult. In this work, we used Brno University of Technology (BUT) phoneme recognizer tool, which is Split-Temporal Context (STC) neural network-based phoneme recognition system [215]. In particular, BUT system trained on the English data from TIMIT corpus have been used to extract PPG. The trained speech recognizers models may not be available for all the languages. Hence, it is necessary to develop the unsupervised speaker-independent posteriorgram.

6.2.2 Gaussian Posteriorgram (GP)

Gaussian posteriorgram has been extensively used for Query-by-Example Spoken Term Detection (QbE-STD) task [216]. In particular, the problem of detecting the presence of a query within the spoken utterance is known as QbE-STD [29, 217]. The posterior probability $P(C_k|\mathbf{o}_t)$ of the current frame $\mathbf{o}_t$ (for k^{th} cluster C_k, and

t^{th} feature vector) of GP can be computed as follows [216]:

$$P(C_k|\mathbf{o}_t) = \frac{\omega^k_{init}\mathcal{N}(\mathbf{o}_t; \mu^k_{init}, \Sigma^k_{init})}{\sum_{j=1}^{N_p} \omega^j_{init}\mathcal{N}(\mathbf{o}_t; \mu^j_{init}, \Sigma^j_{init})}, \tag{6.1}$$

where N_p is the number of GMM components, ω^k_{init}, μ^k_{init}, and Σ^k_{init} are the weights, mean vectors, and covariance matrices, respectively, for each k^{th} Gaussian components ($1 \leq k \leq N_p$). The GMM parameters are estimated using Expectation-Maximization (EM) algorithm [123]. The GMM parameters are trained from a large number of speakers, and the procedure for training the GMM is described in the following sub-Section 6.2.3, step 2.

6.2.3 Proposed VTL-Warped GP

It has been observed that the formants for uniform vocal tract are *inversely* proportional to the length of vocal tract system [17]. Thus, variation in VTL is a wellknown contributing factor to speaker-related spectral variability for having different speaker characteristics. To obtain the speaker-independent posterior features, we applied VTLN technique to achieve speaker-independent VTL-warped posterior features. The conventional approaches for VTLN warping factor estimation require an acoustic model, and a phonetic transcription. The maximum likelihood search is performed to obtain suitable speaker-specific VTLN warping factor [218, 219]. The procedure of VTL-warped Gaussian posteriorgram is as follows [29], [30]:

1. **Feature Extraction**: Compute warped feature sequence, i.e., $\mathbf{X}^\alpha := \{\mathbf{x}^\alpha_1, \mathbf{x}^\alpha_2, \ldots \mathbf{x}^\alpha_T\}$ that carry information from the different warping factors. In this work, we used Perceptual Linear Prediction (PLP) cepstral feature vectors [220]. Human VTL varies from nearly 13 cm (for adult female) to 18 cm (for adult male) [218]. Due to this variations in length, formant frequencies can deviate by 25 % among different speakers. To incorporate this deviation, the VTLN warping factors are considered in the range from 0.88 to 1.12 (i.e., $\alpha = 0.88, 0.90, \ldots, 1.12$) [218].

2. **Initial speaker-independent GMM Training:** The unwarped feature vectors are pulled to model the gender-independent characteristics for acoustic features. The unwarped features, i.e., the feature vector with VTLN warping factor $\alpha = 1$, are taken from a large number of speakers and thus, can be assumed to have speaker-independent characteristics. Let this trained gender-independent GMM be λ having the weight parameters ω_i, mean vectors μ_i,

covariance matrix Σ_i for i^{th} components of GMM. In particular,

$$\lambda_{init} = \{\omega_{init}^i, \mu_{init}^i, \Sigma_{init}^i\}_{i=1}^{N_c}, \tag{6.2}$$

where N_c is the number of components. In a practical or realistic scenarios, covariance matrix Σ_i is considered to have only diagonal elements for computational simplicity. We used this speaker-independent GMM to compute the Gaussian posteriorgram as described in sub-Section 6.2.2.

3. **VTLN factor estimation:** The formants shifting w.r.t. speaker-independent GMM can be captured by VTLN warping factor. In order to estimate the VTLN warping factor, we follow the approach as suggested in our previous work [29], [30]. VTLN warping factor is estimated from the sets of different warped feature vector sequences $\mathbf{X}^{\alpha}$, with different values of α. The likelihood values of $\mathbf{X}^{\alpha}$ are computed against the speaker-independent GMM, λ_{init}, and MLE criteria is considered to estimate VTLN warping factor, i.e.,

$$\hat{\alpha} = \underset{0.88 \leq \alpha \leq 1.12}{\arg\max} \; P(\mathbf{X}^{\alpha} | \lambda_{init}). \tag{6.3}$$

4. **Retraining of GMM:** The VTL-warped features can be combined together from a large number of speakers to train GMM. The objective is to achieve further speaker-independent model by utilizing the acoustic features after spectral scaling compensation. GMM is further retrained with the warped features, X^{α}. This new model, $\lambda_r \sim (\mu_r, \Sigma_r, \omega_r)$ is different from the earlier GMM model, λ_{init}, and expected to have more speaker-independent characteristics.

5. Set $\lambda_{init} = \lambda_r$. Run steps 3 to steps 5 for five times [29], [30].

6. **Computation of Posteriorgram**: Now based on the estimated warping factors and trained GMM, Gaussian posteriorgrams are computed. Thus, the Eq. (6.1) of Gaussian posteriorgram is modified as follows by considering new speaker-independent model, λ_r:

$$P(C_k | \mathbf{o}_t) = \frac{\omega_r^k \mathcal{N}(\mathbf{o}_t; \mu_r^k, \Sigma_r^k)}{\sum_{j=1}^{N_p} \omega_r^j \mathcal{N}(\mathbf{o}_t; \mu_r^j, \Sigma_r^j)}. \tag{6.4}$$

Figure 6.1 shows the block diagram of warped Gaussian posteriorgram computation using VTLN warping factor and speaker-independent GMM. In particular, VTL-warped features are extracted with a different warping factor, α. Initial GMM model is created as speaker-independent GMM using unwarped acoustic features. Speaker-independent GMM is then used to estimate VTLN warping factor for each utterance using MLE criteria. Using estimated VTLN warping factor, VTLN warped acoustic features are used to retrain the GMM parameter to characterize more speaker-invariance property in GMM. This new speaker-independent GMM is further used to estimate VTLN warping factor and warped acoustic features are again used for re-training. This procedure is iteratively executed for five iterations (as suggested in [29], [30]).

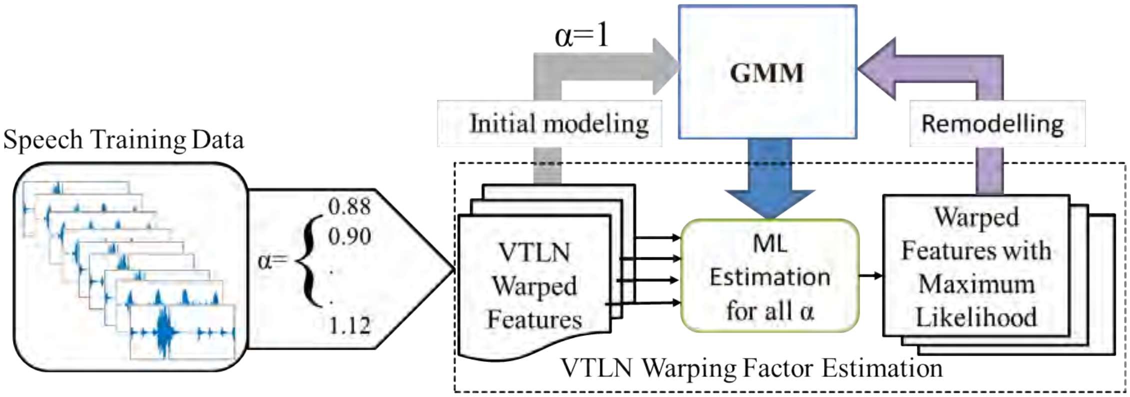


Figure 6.1: Schematic block diagram of extracting VTL-warped posterior features. After [29], [30].

6.2.4 Proposed System Architecture

We employ a two-stage network to develop the VC systems, for both the parallel, and the non-parallel VC (as shown in Figure 6.2). Here, the VTLN warping factor is estimated first for a given speaker-pair. This warping factor is then used to estimate the VTL-warped GP. One DNN (i.e., DNN-A) is trained by taking source speaker's spectral features as an input, and the VTL-warped posteriorgram features as an output. The other DNN (i.e., DNN-B) is trained by considering target speaker's VTL-warped posteriorgram features, and spectral features as an input and output, respectively. During the testing phase, the source speaker's spectral features are first converted using DNN-A, and then this predicted VTL-warped posteriorgram is given as an input to the DNN-B. Converted spectral features predicted by DNN-B is then applied to the vocoder and converted into the speech signal. The proposed framework does not need the time-aligned source and target speakers' features. Thus, the proposed framework behaves exactly the same way for parallel VC as for the non-parallel VC. For DNN-B, we used two architectures, namely, DNN and GAN.

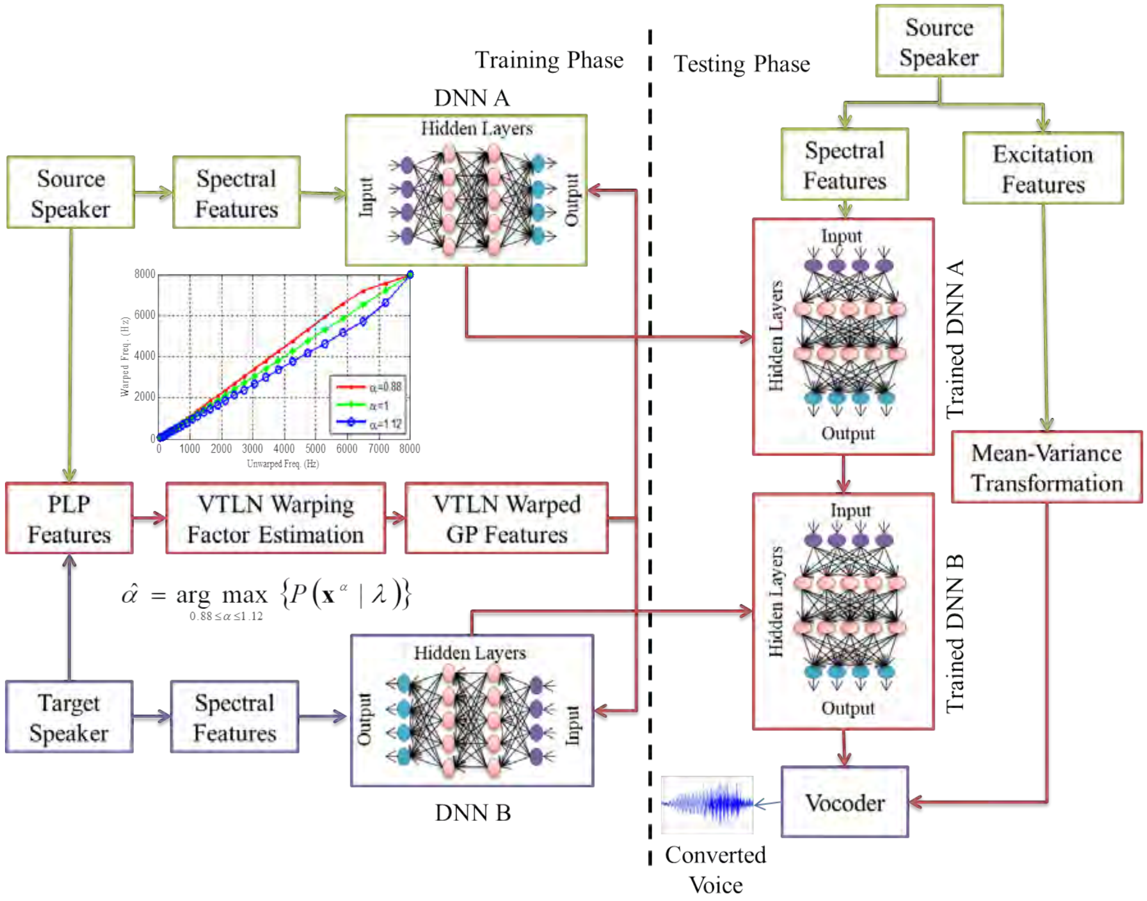


Figure 6.2: The system architecture for proposed VTL-warped GP for non-parallel VC. After [31].

6.2.5 Experimental Results

The experiments are evaluated on the Voice Conversion Challenge 2018 database [119]. The AHOCODER is used for the analysis-synthesis [67]. *40*-dimensional (dim) Mel Cepstral Coefficients (MCCs) (including the 0^{th} coefficient), 39-dim PLP features (including 13-dim static + Δ+ΔΔ), and 1-dim fundamental frequency (i.e., F_0) for each frame (with 25 ms frame duration, and 5 ms frame shift) have been extracted. 120-dim GMM, and VTL-warped posteriorgram are extracted. The number of mixture components were initially set to 128 with Vector Quantization (VQ) initialization. To obtain 120-d posterior features, we perform the iterative approach to merge the two closest centroids till we obtain 120 centroids. Mean-Variance (MV) transformation is used for the F_0 transformation [16].

Two networks were trained to compare the performance of the developed VC systems. The first sub-network that maps the source cepstral features to the VTLN posterior features, remains the DNN with three hidden layers, each having 512 units, and sigmoid as an activation function. The network is trained by minimizing the cross-entropy loss, and softmax activation at the output layer, as we intend to generate the posterior probabilities at the output. The second sub-network that maps the learned VTLN posterior features to the target cepstral features, varies according to the optimization technique adopted. Here, we compare the adversarial loss (MMSE-GAN) w.r.t. the ML-based optimization technique (DNN). The first network is a DNN with MMSE optimization function. In MMSE-GAN, the G network is identical to the DNN. The DNN and G network of the MMSE-GAN has three hidden layers. Each hidden layer with 512 hidden units, is followed by a batch normalization [221], and sigmoid activation function. The output layer has linear activation, so as to generate the target cepstral features. The D network of MMSE-GAN also has three hidden layers, with 512 units followed by batch normalization, and *tanh* activation. The output layer has single unit with sigmoid activation. All the models are trained with 500 epochs, and optimized with an Adam optimizer [140], using a batch size of 1000. All the networks are trained using a learning rate of 0.001.

6.2.5.1 Objective Evaluation

For objective evaluation, we have selected the state-of-the-art Mel Cepstral Distortion (MCD) measure [16]. Table 6.1 shows the average MCD for all the systems developed using 16 speaker-pairs along with the 95 % Confidence Interval (CI) using DNN, and MMSE-GAN for each Hub and Spoke tasks, respectively. The effectiveness of the unsupervised VTLN posterior over a GMM posterior can be easily seen in both the tasks using each method. In particular, we obtained on an average, 0.6 % relative reduction in the MCD with the VTLN posterior compared to the baseline GMM posterior in both the tasks.

Table 6.1: MCD analysis for Hub, and Spoke task. Here, number in bracket indicates a margin of error corresponding to the 95 % CI. After [10]

Posteriorgram	Hub Task		Spoke Task	
	DNN	MMSE GAN	DNN	MMSE GAN
GMM	7.43 (0.16)	7.76 (0.15)	7.63 (0.18)	**7.86** **(0.18)**
Proposed VTLN	**7.36** **(0.16)**	**7.71** **(0.15)**	**7.57** **(0.18)**	**7.86** **(0.18)**

6.2.5.2 Subjective Evaluation

For subjective test, we performed two tests to measure the speech quality, and the Speaker Similarity (SS) of the converted voices from both the tasks. In particular, Mean Opinion Score (MOS) tests have been taken. The subjective tests were taken from the 40 subjects from the total 1920 samples. Subjects were asked to evaluate the randomly played utterances for the speech quality, and SS in MOS evaluations, respectively. In particular, subjects were asked to rate the converted voices on the scale of 1 (i.e., very bad) to 5 (i.e., very good) for speech quality. Similarly, subjects were asked to rate the converted voices in terms of SS on the scale of 1 (totally different) to 5 (exactly similar) w.r.t. the target speaker.

Table 6.2 and Table 6.3 show the analysis of subjective evaluation for the speech quality and speaker similarity, respectively, on both the hub, and spoke tasks, along with their 95 % CI. The effectiveness of VTLN posteriorgram over GMM-based posteriorgram can be seen for the speech quality in both the tasks. We can see from Table 6.3 that the proposed VTLN posteriorgram performs better corresponding to the GMM-based posteriorgram in the case of Hub task. From our analysis, we conclude that the proposed unsupervised VTLN posteriorgram on an average performs better compared to the GMM in speech quality as well as the speaker similarity irrespective of the mapping function used. However, we observed that the GAN-based architectures perform comparable with the two-stage DNN for some of the cases. It may be due to lesser availability of the training utterances in the VCC 2018 database. GAN training requires a large number of samples in minimizing the *distributional* divergence [48,222].

Table 6.2: MOS analysis for speech quality. Here, the number in a round bracket indicates the margin of error related to 95 % CI. After [10]

Posteriorgram	Hub Task		Spoke Task	
	DNN	MMSE-GAN	DNN	MMSE-GAN
GMM	3.15 (0.18)	2.57 (0.16)	**2.87** **(0.15)**	**2.82** **(0.16)**
Proposed VTLN	**3.32** **(0.16)**	**3.1** **(0.14)**	2.08 (0.14)	2.62 (0.17)

In this study, we have presented a novel unsupervised VTLN posterior representation for parallel as well as non-parallel VC task. In particular, a two-stage voice conversion system, wherein a DNN, and MMSE-GAN have been used as a synthesizer, and evaluated on the VCC 2018 database. From both objective as well as subjective evaluation, we observed that the proposed unsupervised VTLN-based posterior features performs on an average better over the GMM-based posterior features in all the mapping function cases in both parallel as well as non-parallel VC.

Table 6.3: MOS analysis for speaker similarity. Here, number in a bracket indicates the margin of error related to 95 % CI. After [10]

Posteriorgram	Hub Task		Spoke Task	
	DNN	MMSE-GAN	DNN	MMSE-GAN
GMM	2.54 (0.24)	2.13 (0.28)	**2.72** **(0.24)**	**2.5** **(0.22)**
Proposed VTLN	**3.05** **(0.27)**	**2.86** **(0.24)**	2.69 (0.27)	2.19 (0.25)

6.3 Novel Inter Mixture Weighted (IMW) GMM Posteriorgram

6.3.1 Limitations of GMM posteriorgram

The key issue with the conventional GMM-PG is that the same phone gets spread across more than one component (due to the speaking style variations across the speakers) [217]. In particular, we observe that this spread is limited to a group of neighboring components for a given phone. In this work, entire TIMIT database is used to train the GMM. In particular, we have analyzed frame-level GMM-PG features for three randomly selected speakers from the TIMIT database. For selected phones, we present in Table 6.4 that the indices of those components, that are having higher probabilities in the decreasing order from the top, for a specific speaker. An ideal posterior speaker-independent representation should contain distinct phonetic information in each component, irrespective of the speaker. However, in the case of GMM-PG, the phonetic information gets spread across the components.

From Table 6.4, it can be observed that the components that share the frame posteriorgram values are almost the same for one user and for one phone. In case of phones, for example, 'aa ', the found component labels $46, 53, 26, 40$ are representing one particular phone, across the speakers. However, while observing the parameters (i.e., *mean* and *variance*) of these components, we observed the distance calculated by KL-divergence is lesser w.r.t. each other. This means that the feature vectors for a phone lies closer across the speakers even though they cannot be clustered to the same component. Furthermore, it should be noted that across different speakers (as shown in Table 6.4), prominent (in terms of having highest posterior probability) components remains similar for most of the cases. Though they are similar, their order of prominence varies. This explains why the prominent component of a phone, for example, $/iy/$ is 25^{th} component for *male* speaker, while for *female* speaker, it is 29^{th} component. While using conventional GMM-PG matching methods, this may lead to many false recognition or increased mismatch [217]. To address this issue, we propose Inter Mixture Weighted GMM-PG (i.e., IMW GMM-PG), that shares the posterior probability of each mixture in GMM-PG with the limited number of neighboring mixtures that are *sorted* based on the Kullback-Leibler (KL) divergence measure [11].

Table 6.4: Prominent component variation in GMM posteriorgram, 64 dimensions, for selected phones, and speakers from TIMIT database. After [11]

Speaker	Selected Phonetic Classes									
	aa	**iy**	**ow**	**l**	**m**	**dh**	**f**	**sh**	**p**	**t**
Male1	46	25	44	51	3	54	14	1	7	15
	53	58	56	44	19	63	35	35	15	28
	26	29	51	3	23	31	1	9	9	7
Female1	46	29	37	44	3	15	13	14	4	7
	26	22	18	32	23	39	35	35	9	4
	40	25	41	20	19	63	4	4	15	9

6.3.2 Proposed IMW GMM Posteriorgram

IMW GMM is a post-processing method obtained by sorting components based on the distance calculated among the components of GMM using KL divergence measure. The neighboring components of a given component, are assigned a fraction of its probability value along with their posterior probability value. This in effect helps feature being represented by a set of components than a single component. The block diagram to extract IMW GMM-PG is shown in Figure 6.3. Corresponding algorithm is shown in Algorithm 3. The details of the IMW GMM-PG extraction are given in the following Algorithm 3.

Algorithm 3 Proposed IMW GMM Algorithm. After [11]

extractIMWGMMPost()

extractIMWGMMPost returns p' IMW GMM-PG which is obtained from GMM-PG p after applying IMW GMM logic.

```
Input: $N_c \leftarrow$ Number of components in GMM.
       $n \leftarrow$ Dimension of a feature vector.
       $\theta \leftarrow$ Model parameters $\mu, \sigma, \omega$.
       $p(i) \leftarrow$ Posterior probability of a $i^{th}$ component for a frame calculated using GMM ($1 \times N_c$).
$S \leftarrow$ IMWGMM(n, $N_c$, $\theta$)
$k \leftarrow \arg\max_{i} p(i)$, index of component
          with maximum probability in $p$.
$i \leftarrow 1$
while $i \leq N_c$:
    $M \leftarrow$ position of i in $S[k,:]$. (To find how closer $i^{th}$ component is from $k$)
    $p'(i) \leftarrow p(i) + \frac{p(k)}{2^M}$
end
return $p' \leftarrow$ normalize $p'$
```

function *IMWGMM(n,N,θ)*

IMWGMM returns a matrix S, which includes the indices of components sorted in ascending order of KL distance from each component.

```
$i \leftarrow 1$
$S \leftarrow$ zero initialized matrix, of dimension $N_c \times N_c$,
    (Stores indices of nearest component, $\forall\, i$)
$D \leftarrow$ zero initialized array of dimension, $1 \times N$
    (Stores distance array of a component.)
while i $\leq N_c$:
    $D \leftarrow$ calcDist(i θ)
    $i \leftarrow i + 1$
    $S[i,:] \leftarrow$ Indices of D after sorting in ascending order.
end
return $S$
```

function *calcDist(i,θ)*

calcDist returns array d : KL distance of i^{th} component from all N components given by θ.

```
Input: $i \leftarrow$ Component under consideration
for $j$ from 1 to $N$ :
    $P \leftarrow \theta(i)$
    $Q \leftarrow \theta(j)$
    $d[j] \leftarrow D_{KL}(P||Q)$
end
$d[i] \leftarrow$ high value, to avoid the same component to be detected as neighbor
return $d$
```

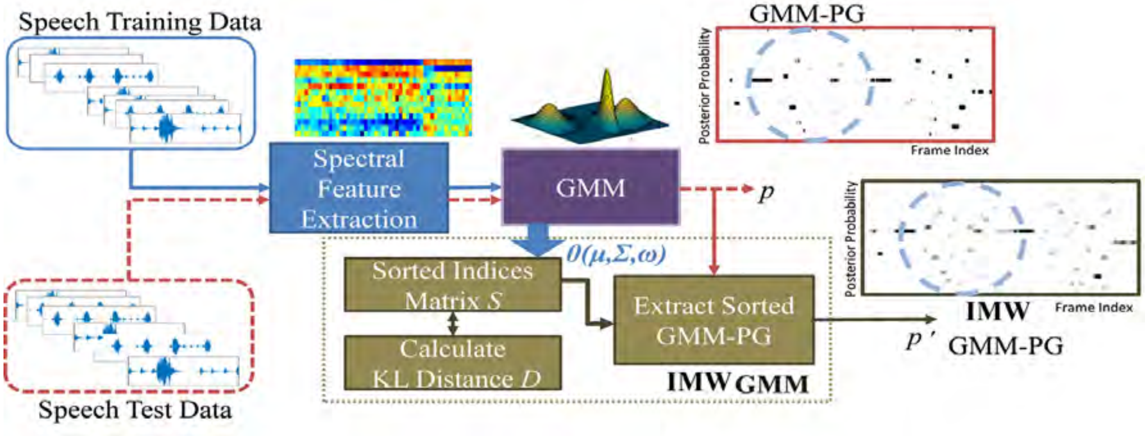


Figure 6.3: Functional block diagram of proposed IMW GMM method for VC. It can be inferred from the circled regions that the probability is shared across components in IMW GMM-PG than in GMM-PG. After [11].

After extracting GMM-PG for a frame, the component having the highest probability is selected, and its neighbor components are identified in a sorting order. Depending on the order, the probability value is shared with all the components, in such a way that the nearest neighbor components get the highest share, while the farthest components gets the lowest share. This results in spreading the posterior probability across the neighbor components which can be seen in the circled region in Figure 6.4.

To visualize the effectiveness of IMW GMM-PG over GMM-PG, we calculated the distance among posterior features across different phonetic classes on entire TIMIT database. The broad phone classes considered here includes vowels, semi vowels, nasals, voiced fricatives, unvoiced fricatives and plosives.

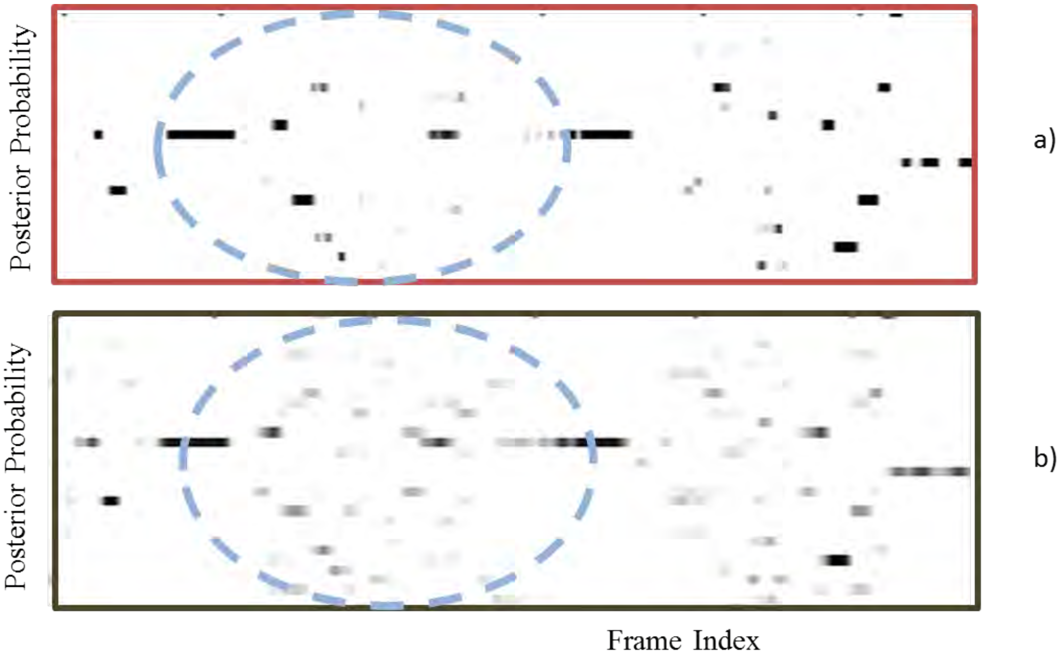


Figure 6.4: Posteriorgram comparison for the query, *"intelligence"*, (a) GMM-PG, and (b) IMW GMM-PG. After [11].

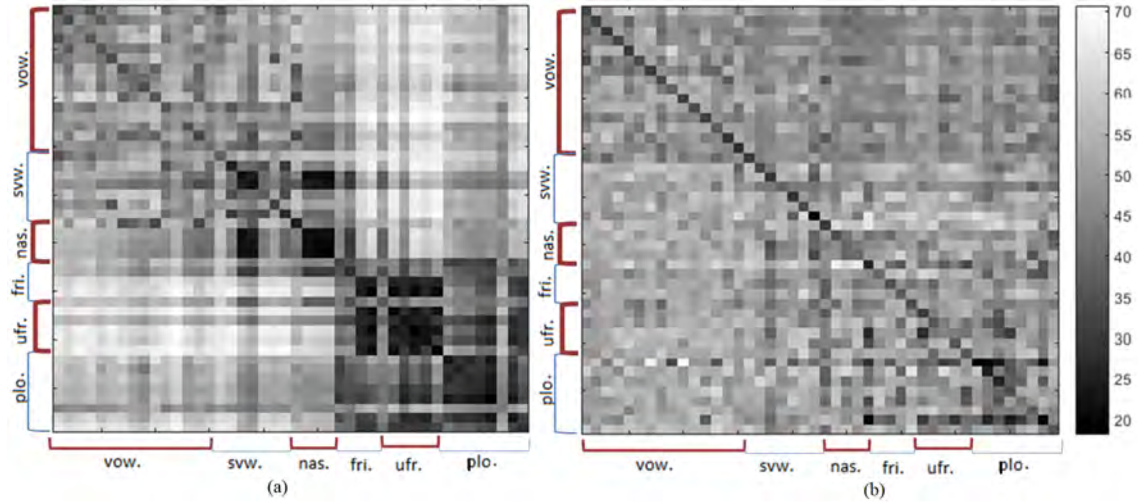


Figure 6.5: Distance matrix with 44 phones for (a) GMM, and (b) proposed IMW GMM approach. Here, vow: vowels, svw: semi vowels, nas: nasals, fri: voiced fricatives, ufr: unvoiced fricatives, and plo: plosives. After [11].

Figure 6.5 clearly shows that in the case of obstruents (i.e., fricatives, affricates, and plosives), there is more ambiguity with the GMM-PG features than the IMW GMM-PG features. Vowels and semivowels broad classes are easily distinguished within in case of IMW GMM-PG. However, it can also be noticed that in GMM-PG the vowel and semivowels can be very clearly distinguished from the plosive, fricatives broad phone classes, while the distinction between the broad classes decreased in case of IMW GMM-PG.

6.3.3 Proposed VC System Architecture

We develop the two-stage DNN-based VC architecture. In particular, the first stage maps the cepstral features of a source speaker to the corresponding source speaker's IMW GMM-PGs using DNN, and the second network maps the target speaker's IMW GMM-PGs to the corresponding cepstral features of the target speaker using DNN and GAN (as shown in Figure 6.6 and Figure 6.7).

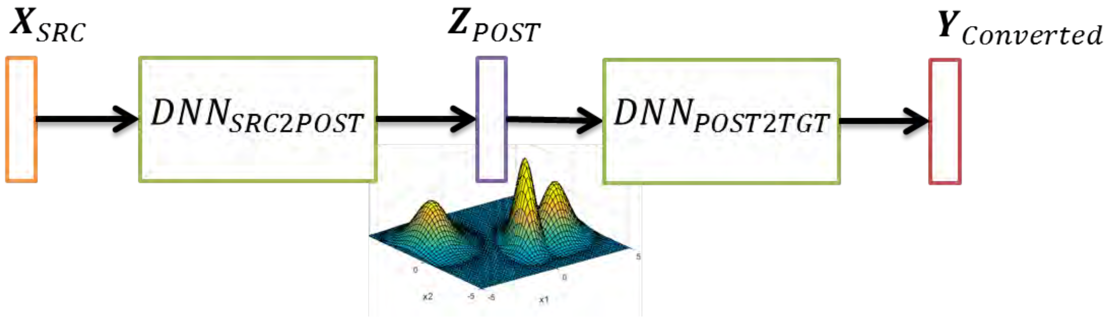


Figure 6.6: Schematic representation of the proposed two-stage DNN VC framework. After [11].

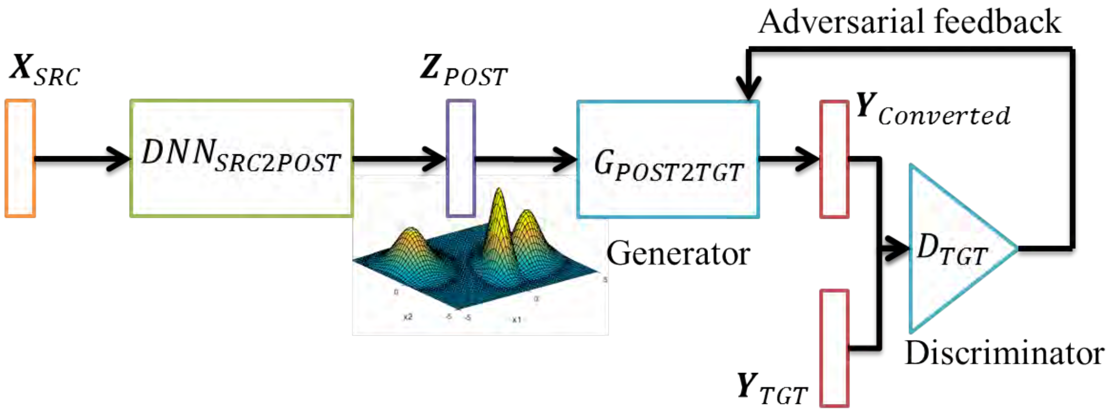


Figure 6.7: Schematic representation of the proposed hybrid DNN-GAN VC framework. After [11].

Both the networks are trained simultaneously. During the time of conversion, the posterior features are predicted from the source speaker's cepstral features using the first network (i.e., $DNN_{SRC2POST}$). These predicted posterior features are then passed through the second network to predict the target speaker's cepstral features (i.e., $DNN_{POST2TGT}$). Since the GAN is able to produce natural realistic samples, we also propose to use GAN at the second stage for synthesis of the target cepstral features from the posterior features, instead of a simple DNN-based network. The block diagram of the GAN-based network is shown in Figure 6.7.

6.3.4 Experimental Results

The VCC 2016 database consists of training utterances from 5 source and 5 target speakers [118]. In this work, we develop *25* VC systems using each method among the available speaker-pairs. AHOCODER have been used for analysis-synthesis framework [67]. We extract 25-dimensional (dim) Mel Cepstral Coefficient (MCC) features over a 25 ms window duration with the 5 ms frame shift. For each speaker in VCC 2016 database, *64*-dim GMM and SGMM posteriors are extracted based on the model trained on the TIMIT database on *39*-dim MFCC (including *13*-dim static + Δ + $\Delta\Delta$ features).

The G network in the MMSE-GAN has three hidden layers, with 512 hidden units. Each layer is followed by batch normalization [221], and sigmoid activation. The output layer has 25 units to predict the target cepstral features, with linear activation. The D network also has three hidden layers, with 512 hidden units, and each followed by batch normalization, and *tanh* activation. The last layer uses the sigmoid activation in the D network. Dropout with 0.3 drop probability is selected for all the hidden layers in the G and D networks. The network is trained for 250 epochs, with an effective batch size of 1000. The network parameters are updated through Adam optimization [140], with a suitable learning rate of 0.001 [48]. Once the network is trained, the model with the least Minimum Square Error (MSE) on the validation set is selected, and the testing is performed.

6.3.4.1 Subjective Evaluation

In this work, two Mean Opinion Score (MOS) tests have been performed to evaluate the developed VC systems, based on the speech quality, and the Speaker Similarity (SS) of the converted voices. 26 subjects (4 females and 22 males without any hearing impairments, and with the age variations between 18 to 22 years) participated in both the tests. Subjects evaluated the randomly played utterances for the speech quality on *5*-point scale. Figure 6.8 shows the MOS analysis (obtained from total 384 samples) for the developed VC systems along with their 95 % confidence interval to quote the statistical significance of the results. Effectiveness of the proposed IMW GMM-PG over GMM-PG is ubiquitous in both the architecture in the context of speech quality of the converted voices. In particular, we obtained on an average 19.52 %, and 7.94 % relative improvement in MOS for speech quality with the DNN, and the GAN-based VC systems, respectively.

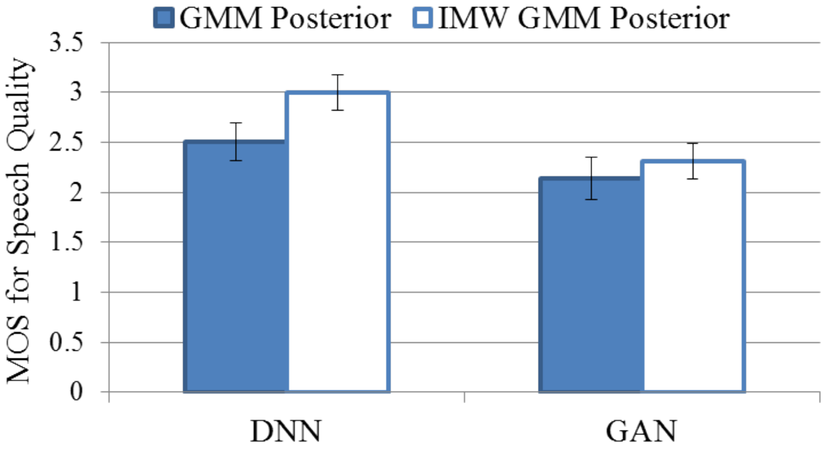


Figure 6.8: MOS scores w.r.t. the speech quality of the developed systems along with the 95 % confidence interval. After [11].

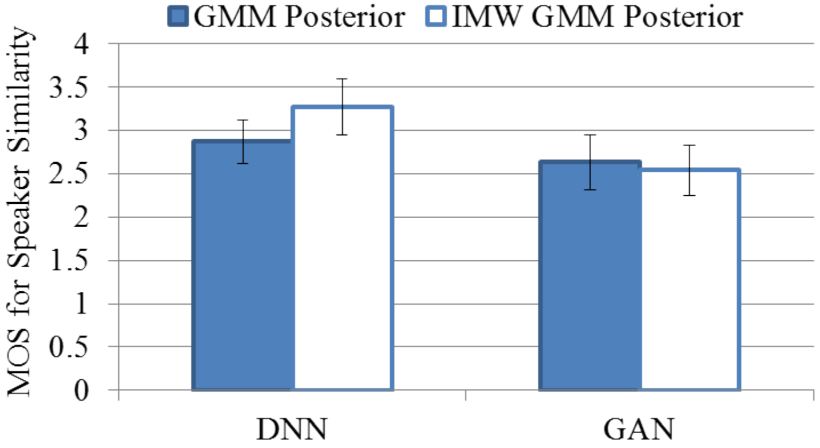


Figure 6.9: MOS scores w.r.t. the speaker similarity (SS) of the developed systems along with the 95 % confidence interval. After [11].

Similarly, in another MOS test, on a 5-point scale subjects rated the converted voices in terms of SS w.r.t. the target speaker. In the 5-point scale, 1 means totally different to the target speaker, and 5 means exactly similar to the target speaker for the SS. Figure 6.9 shows the MOS for the SS of the developed VC systems along with their 95 % confidence interval. In particular, we obtained on an average 5.25 % of relative improvement in terms of MOS for SS with the proposed IMW GMM-PG features compared to the GMM-PG features. The lack of a large number of training examples for the adversarial training, results in lower performance of the GAN w.r.t. the DNN-based VC system for speech quality and SS, respectively (as shown in Figure 6.8 and Figure 6.9). However, the proposed IMW GMM-PG features clearly outperform the conventional GMM-PG.

6.3.4.2 Objective Evaluation

In this work, traditional objective measure, namely, Mel Cepstral Distortion (MCD) (in dB) has been used for the objective evaluation [16]. The systems having lower MCD can be considered as the better compared to the system having higher MCD. We obtain 0.2 dB of absolute reduction in the MCD with the proposed IMW GMM-PG w.r.t. to the GMM-PG in the DNN-based VC systems as shown in Figure 6.10. The MCD computes the scores on the basis of the numerical similarity between the cepstral features corresponding to the converted, and the target speaker's data. However, the adversarial optimization minimizes the distributional divergence, and do not optimize the numerical difference between the converted, and the target speakers' cepstral features. Hence, we observe increment in the MCD in the case of GAN-based VC systems.

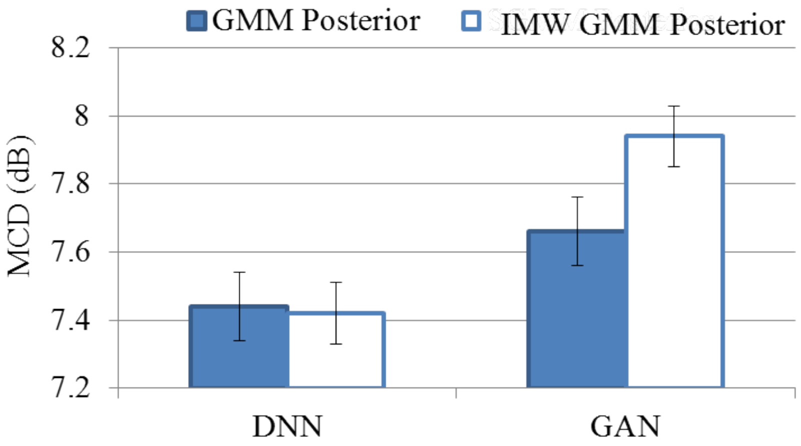


Figure 6.10: MCD scores of the developed systems along with the 95 % confidence interval. After [11].

6.4 Chapter Summary

In this chapter, we proposed two-stage conversion strategy in order to avoid alignment step in both the VC tasks. The direct alignment procedure between the source and target spectrum results into many-to-one or one-to-many correspondences, which deteriorate the performance of the VC system. This is achieved by training two separate DNNs, where one DNN maps source speaker's spectral features to the speaker-independent representations, and the another DNN maps these speaker-independent representations to the particular target speaker's spectral features. In particular, we propose novel unsupervised VTLN posterior features, and IMW GMM-PG features as the speaker-independent representations to overcome the need of alignment in the case of two-stage DNN as well as GAN-based VC frameworks.

The speaker-independent VTLN posterior feature set is obtained by training a GMM with vocal tract length (VTL) warped cepstral features extracted from the data of the pair speakers. The warping factor is found iteratively with GMM training such that maximum likelihood is obtained under a discrete set of warped values and given the speaker-pair non-parallel acoustic data. The final VTLN posterior feature set is the Gaussian posteriograms computed from the resulting GMM. In addition, we also proposed IMW GMM-PG as an unsupervised speaker independent representation in the proposed two-stage mapping network. The key idea of IMW GMM-PG feature is to share the probability values of the current component with its neighbor, to spread the posterior probability across the components. The effectiveness of the proposed IMW GMM-PG features over the GMM-PG features can be clearly observed in the context of VC tasks. In the next chapter, we will discuss the issues associated with the objective evaluation of the VC systems and propose novel application of Acoustic-to-Articulatory Inversion (AAI) for quality assessment of converted voices.

CHAPTER 7

Quality Assessment of VC

7.1 Introduction

In this chapter, we propose a novel application of the Acoustic-To-Articulatory Inversion (AAI) towards a quality assessment of the voice converted speech. The ability of humans to speak effortlessly requires the coordinated movements of various articulators, muscles, etc. This effortless movement contributes towards a naturalness, intelligibility, and speaker's identity (which is partially present in voice converted speech). During VC, some of the important details in the speech signal are lost due to inaccurate spectral mapping and statistical averaging (i.e., oversmoothing) of speech sound units. Investigating the evaluation measure that truly quantifies the naturalness and speaker similarity of a voice converted speech is still an open research problem [15]. Subjective measures are time-consuming, expensive, and their accuracy highly depends on the *cognitive* factors (such as alertness) of the listeners [223]. Objective measures, on the other hand, often lack the intuitiveness as well as do not account for the perceptual quality [224].

Machine generated speech, i.e., any computational way of producing the speech signal can never match the way humans articulate to produce speech [225,226]. In addition, the quality and intelligibility of a voice converted speech are governed mainly by the accurate production of vowels, dynamic or transitional sounds (such as diphthongs, liquids, glides, and stops) [227]. Thus, the study of an articulatory parameters (those which are critical in the production of these speech sounds) could be useful in the voice quality measurement [228–230]. This idea motivated the author to investigate the difference between a voice converted speech, and a natural speech in terms of articulatory parameters. To the best of authors' knowledge, this is in contrast to the previous objective measures which measure the quality in terms of information loss in the spectral characteristics during VC [16,231,232]. The effectiveness of articulatory features has been shown in various applications, such as visual aids for training speech [233], speaker recognition [234], speech recognition [235], accent conversion [236], etc. The VC is an another application where the possibility of using articulatory parameters has been explored. However, it appeared that the use of articulatory parameters was not straightforward for improving VC [237]. In this chapter, we investigate the novel application of articulatory features for the quality assessment of a voice converted speech. This study investigates the following questions:

1. Whether the articulatory information is lost during the VC process?
2. How can one quantify the information loss?

To address this, we propose a novel Estimation Error (EE), an articulatory features-based objective measure.

7.2 Proposed Objective Measure for Quality Assessment

This Section briefly discusses the state-of-the-art techniques, which are used to develop VC and AAI systems.

7.2.1 MOCHA Database

The Multichannel Articulatory (MOCHA) database consists of a simultaneously recorded (460 phonetically diverse British English TIMIT sentences) acoustic, and articulatory data obtained from one male and one female speaker [238]. The audio signal is sampled at 16 kHz and Electromagnetic Articulography (EMA) data is sampled at 500 Hz. The EMA data consists of X and Y coordinates of 9 receiver sensor coils attached to 9 points along the midsaggital plane, namely, the lower incisor or the jaw (li_x, li_y), upper lip (ul_x, ul_y), lower lip (ll_x, ll_y), tongue tip (tt_x, tt_y), tongue body (tb_x, tb_y), tongue dorsum (td_x, td_y), velum (v_x, v_y), upper incisor (ui_x, ui_y), and bridge of the nose (bn_x, bn_y). The upper incisor and bridge of the nose are used as a reference coils. The articulatory data obtained from 14 channels corresponding to first seven coils except the reference coils are used as the articulatory features in our experiments.

7.2.2 Details of AAI System

Among the various available AAI techniques [233, 239, 240] here, Generalized Smoothness Criterion (GSC)-based, AAI system is used for the articulatory parameterization of a voice converted speech [240]. The estimated trajectories obtained using GSC were optimal in the sense that a) the estimated trajectories have minimum energy in the high-frequency region, and b) the weighted difference between estimated and original trajectories was minimum. GSC has the advantage that it imposes the articulator-specific constraints, which gives a better estimation over methods using a fixed smoothness constraints [240].

7.2.3 Proposed Objective Measure

The experiments were conducted to verify, and quantify the possible loss of an articulatory information after VC. For this, a GMM-based VC system with 400 training utterances, and 64 mixture components was used. Let the target and the voice converted acoustic vector be given by $\mathbf{X}_t$ and $\mathbf{X}_{tv}$, respectively. Furthermore, let EMA vector of the target be $\mathbf{Y}_t$ and estimated EMA vector from $\mathbf{X}_t$ and $\mathbf{X}_{tv}$ be $\mathbf{Z}_t$ and $\mathbf{Z}_{tv}$, respectively. In order to verify the loss in speech production information after VC, mutual information (I) was computed [241]. Since $\mathbf{X}_t$, $\mathbf{Y}_t$ and $\mathbf{X}_{tv}$ are discrete, their probability distributions are calculated by quantizing the acoustic and the articulatory spaces using K-means clustering algorithm $(K = 64)$ [242]. Mutual Information (I) calculated between $(Q(\mathbf{X}_t), Q(\mathbf{Y}_t))$ and $(Q(\mathbf{X}_{tv}), Q(\mathbf{Y}_t))$ is shown in Table 7.1. Here, $Q(\mathbf{X}_t)$ and $Q(\mathbf{Y}_t)$ are quantized acoustic and articulatory spaces, respectively, and $Q(\mathbf{X}_{tv})$ is quantized voice converted acoustic space. Table 7.1 shows that the information related to the articulators in acoustic vector reduces after VC both for male (i.e., the target is a male), and female (i.e., the target is a female) voice converted speech.

Table 7.1: Comparison of Mutual Information (MI) before, and after VC. After [12]

I (in bits)	Male Voice	Female Voice
$I(Q(\mathbf{X}_t), \mathbf{Y}_t))$	1.402	1.502
$I(Q(\mathbf{X}_{tv}), \mathbf{Y}_t))$	**1.28**	**1.389**

The following steps were used to estimate the articulatory parameters of a voice converted speech (which is illustrated in Figure 7.1) in order to quantify above mentioned loss.

- $\mathbf{Z}_{tv}$ and $\mathbf{Z}_t$ were estimated using GSC-based technique.
- $\mathbf{Z}_{tv}$, $\mathbf{Z}_t$, and $\mathbf{Y}_t$ were time-normalized (by applying DTW on $\mathbf{X}_{tv}$ and $\mathbf{X}_t$) to obtain $\mathbf{DZ}_{tv}$, $\mathbf{DZ}_t$, and $\mathbf{DY}_t$, respectively.

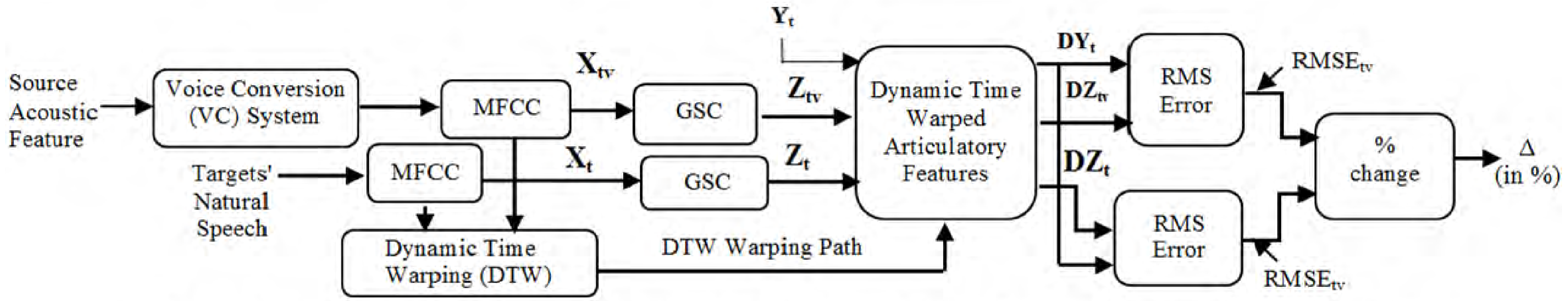


Figure 7.1: Proposed system architecture for estimating articulatory features from VC system. After [12].

- The estimation accuracy for each articulator position was compared by computing % Δ given by:

$$\% \text{ Change}(\Delta) = \frac{RMSE_{tv} - RMSE_{tt}}{RMSE_{tt}} \times 100, \tag{7.1}$$

  where $RMSE_{tt}$ is the average RMSE calculated between $\mathbf{DY}_t$ and $\mathbf{DZ}_t$, and $RMSE_{tv}$ is an average RMSE calculated between $\mathbf{DY}_t$ and $\mathbf{DZ}_{tv}$.

Table 7.2 shows that $RMSE_{tv} > RMSE_{tt}$ for both male, and female voice converted speeches, which is indicated by positive % Δ for all the articulators. In particular, among all the articulators, tongue tip (known to be critical for the speech production [226]) shows the highest % Δ. The results indicate that the AAI system poorly estimates the articulatory trajectories of a voice converted speech. The difference in the estimation accuracy is utilized to propose the Estimation Error (EE), as an objective measure. The EE measures the distance between articulatory trajectories of a voice converted speech, and the target speech. Estimation Error (EE) (in *mm*), is defined as [12]:

$$EE = \frac{1}{N}\Big(\sum_{n=1}^{N}\sqrt{\sum_{d=1}^{M}\Big(\mathbf{DZ}_{tv_d}^{n} - \mathbf{DY}_{t_d}^{n}\Big)^2}\Big), \tag{7.2}$$

where for n^{th} frame, $\mathbf{DY}_{t_d}^{n}$, and $\mathbf{DZ}_{tv_d}^{n}$ are the time-aligned $d^{th}-$dimensional measured, and estimated trajectory, respectively. In addition, N is the length, and M is the dimensionality of the articulator trajectory.

7.3 Experimental Results

7.3.1 Details of VC and AAI systems

Here, the VC systems based on GMM, and BLFW+AS were built for both M-F and female-to-male (F-M) cases. For this, the number of training utterances (i.e., 10, 25, 50, 200, and 400), and the number of mixtures in GMM (i.e., m=8, 16, 32, 64) were varied. 24-D Mel Cepstral Coefficients (MCC) were extracted from the speech signals over 25 ms window with 5 ms shift for both the VC approaches. The training sentences were selected based on maximum diphone coverage [16]. For AAI, out of 400 (from 460 the MOCHA-TIMIT) sentences used for training of VC system, 368 sentences for the development set, and 55 for test set were used. *14*-D MFCC was calculated per frame (of 20 ms window with a frame shift of

Table 7.2: Comparison of an average *RMSE* in *mm*. Here, number in the round bracket indicated the Standard Deviation (SD). After [12]

Voice	Articulators	li_x	li_y	ul_x	ul_y	ll_x	ll_y	tt_x	tt_y	tb_x	tb_y	td_x	td_y	v_x	v_y
Male	$RMSE_{tt}$	0.6	1.11	0.77	1.36	1.28	2.09	2.83	**3.5**	2.66	2.56	2.35	2.67	0.56	1.19
	(SD)	(0.1)	(0.3)	(0.2)	(0.2)	(0.3)	(0.4)	(0.8)	**(0.8)**	(0.6)	(0.5)	(0.6)	(0.5)	(0.2)	(0.5)
	$RMSE_{tv}$	0.63	1.21	0.81	1.47	1.39	2.35	3.19	**3.87**	2.91	2.86	2.58	2.94	0.62	1.29
	(SD)	(0.1)	(0.2)	(0.2)	(0.3)	(0.3)	(0.4)	(1)	**(0.7)**	(0.7)	(0.6)	(0.7)	(0.6)	(0.2)	(0.5)
	% Δ	5	9	5.2	8.1	8.6	12.4	**12.7**	10.6	9.4	11.7	9.8	10.1	10.7	8.4
Female	$RMSE_{tt}$	0.87	1.36	1.01	1.36	1.32	**2.92**	2.72	2.89	2.49	2.61	2.29	2.7	0.45	0.49
	(SD)	(0.2)	(0.3)	(0.4)	(0.3)	(0.3)	**(0.6)**	(0.6)	(0.6)	(0.5)	(0.5)	(0.5)	(0.5)	(0.2)	(0.2)
	$RMSE_{tv}$	0.93	1.5	1.1	1.41	1.42	3.22	3.2	**3.36**	2.88	2.99	2.61	2.94	0.52	0.54
	(SD)	(0.2)	(0.3)	(0.4)	(0.3)	(0.3)	(0.7)	(0.7)	**(0.6)**	(0.6)	(0.5)	(0.6)	(0.4)	(0.2)	(0.2)
	% Δ	6.9	10.3	8.9	3.7	7.6	10.3	**17.6**	16.3	15.7	14.6	14	8.9	15.6	10.2

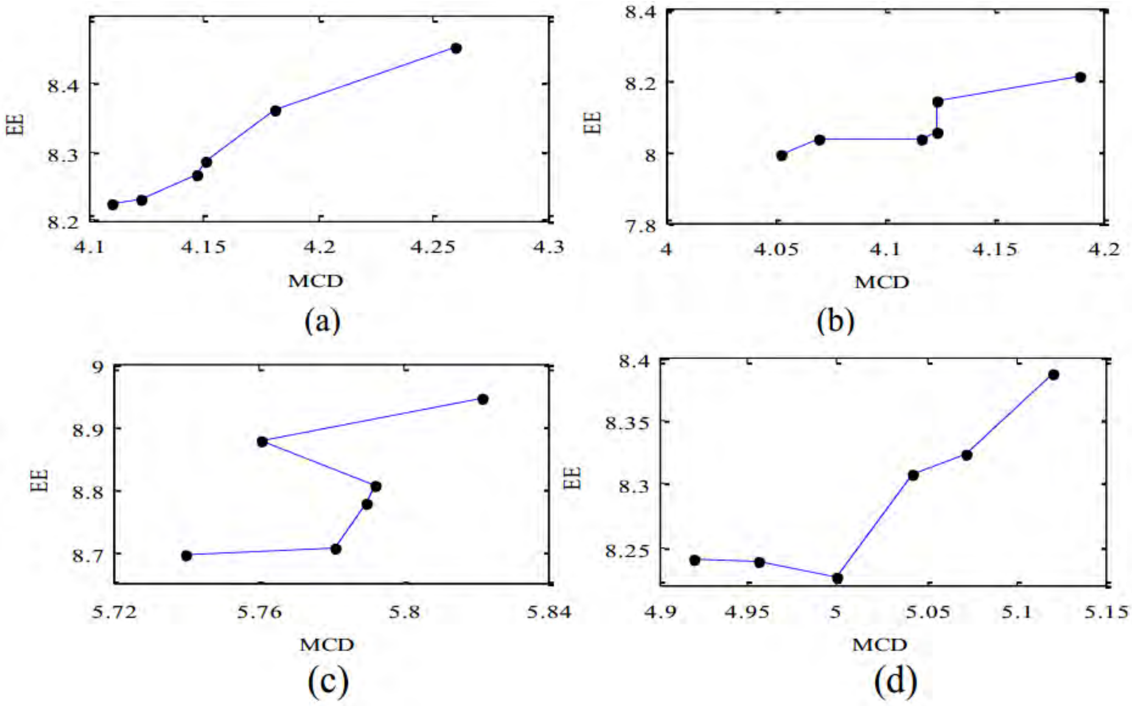


Figure 7.2: MCD *vs.* plot for selected systems (a)-(b) M-F and F-M GMM-based VC, and (c)-(d) M-F and F-M BLFW+AS-based VC. After [12].

10 ms) for inversion. AAI systems were built for both male and female voices. The accuracy of AAI system is measured by calculating an average RMSE and an average Correlation Coefficient (CC) [240]. Our AAI system shows the lowest estimation accuracy for ll_y (average $RMSE = 2.92$, average $CC = 0.74$), and highest for v_x (average $RMSE = 0.45$, average $CC = 0.70$) in case of a female. For a male, the estimation accuracy is the lowest for tt_y (average $RMSE = 3.5$, average $CC = 0.70$), and highest for v_x (average $RMSE = 0.56$, average $CC = 0.64$).

7.3.2 Correlation of EE with objective measure

Figure 7.2 shows plot between EE, and MCD for the selected systems. These plots indicate that EE and MCD are partially correlated. In particular, Figure 7.2 (a)-(b) show that EE and MCD correlate well for GMM–based VC as compared to the BLFW+AS-based VC (as shown in Figure 7.2 (c)-(d)). One of the possible reasons for such a high correlation could be that articulatory parameters were estimated from the acoustic features itself. However, the two speech sounds that are closer in cepstral or acoustic-domain may not be close in articulatory-domain, because AAI is a non-unique, and nonlinear [233, 240]. Moreover, it is known that MCD may not always correlate well with the subjective scores [5, 14, 45, 49]. Therefore, the differences in articulatory-domain were exploited for determining the quality of the VC systems. To verify this, the Pearson Correlation Coefficient (PCC) of MCD, EE with subjective measures were calculated.

7.3.3 Comparison of EE with the subjective measures

For a subjective measure, MOS from 15 subjects (9 male and 6 female with age group of 21-25 years and no known hearing impairments) was taken for absolute rating [171]. In this test, we randomly played 4 sentences from each VC system (selected for evaluation). The subjects were asked to score them on a 5-point scale based on the naturalness of speech signal. CC of MCD and EE with MOS score was calculated using the Pearson Correlation Coefficient (PCC).

Table 7.3: Subjective and objective scores of various VC systems. After [12]

Approach	Systems*	M-F VC			F-M VC		
		MOS	MCD	EE	MOS	MCD	EE
BLFW+AS	10_64	2.45	5.66	7.6	2.35	4.87	8.05
	25_64	2.65	5.65	7.68	2.45	4.84	7.72
	50_64	2.53	5.71	7.59	2.33	4.97	7.9
	100_64	2.63	5.99	7.96	2.68	5.36	8
	200_64	2.4	6.09	8.17	2.63	5.26	8.29
	400_64	2.33	5.89	8.11	2.6	5.12	8.03
GMM	10_32	2.48	3.97	7.76	2.1	3.98	7.28
	25_32	2.3	4.04	7.29	2.2	3.92	6.92
	50_64	2.53	3.8	7.42	2.15	3.93	7.12
	100_64	2.53	4.24	7.61	2.18	4.16	7.03
	200_64	2.23	4.08	7.76	2.3	4.09	7.36
	400_64	2.35	4.235	7.438	2.225	4.09	7.04

*Systems: Number of Training Utterances_Mixture Components

MOS score, MCD, and EE for selected systems along with their CC are shown in Table 7.3, and Table 7.4, respectively. Ideally, subjective and objective scores should have a negative correlation. Since higher the MOS better is the quality of speech, as opposed to MCD and EE where higher the score lesser is the quality. Table 7.4 shows that for M-F MCD and EE are showing negative CC for both VC techniques. It can be seen from Table 7.4 that EE is more *negatively* correlated with MOS in the case of M-F. On the other hand, in the case of F-M, EE is relatively less positively correlated compared to the MCD. Hence, it is found that the interpretation of a quality given by EE was more preferable over MCD measure. While conducting the MOS test, it was observed that there were minute perceptual differences within VC systems used for evaluation and subjects had difficulty in giving MOS scores. This can be observed from Table 7.3, which indicates a very small change in MOS scores.

Table 7.4: PCC of MCD and EE with MOS. After [12]

Objective Measure	GMM		BLFW+AS	
	M-F	F-M	M-F	F-M
MCD	-0.16	0.41	-0.33	0.87
EE	-0.7	0.16	-0.5	0.46

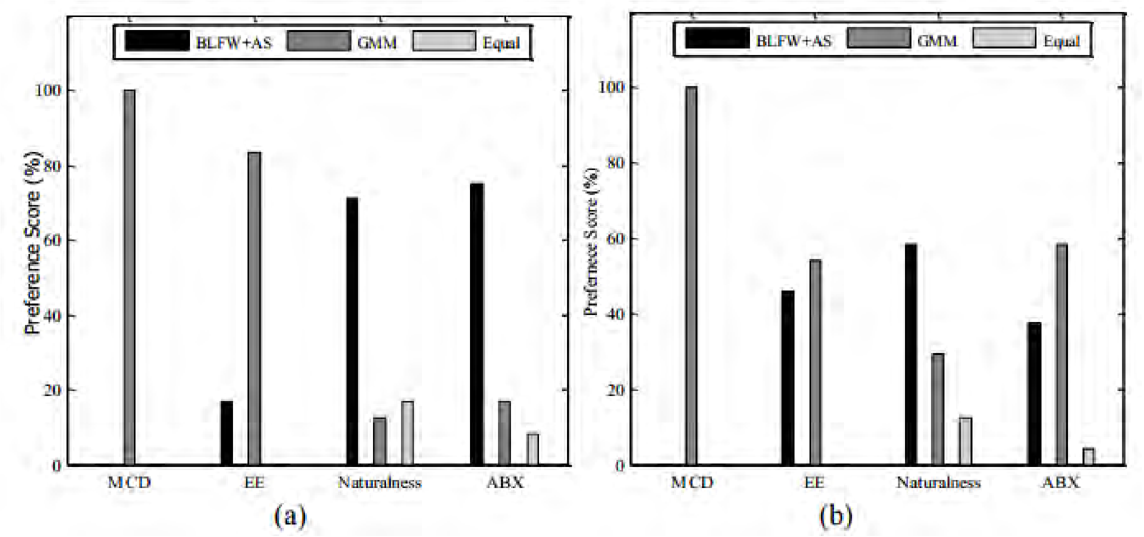


Figure 7.3: Preference score based on MCD, EE, naturalness, and ABX test for GMM, and BLFW VC systems (a) M-F (b) F-M. Equal means, subjects could not judge and give equal preference score. After [12].

In order to avoid this ambiguity, ABX test was conducted with the same 15 subjects, where 24 utterances were played randomly from both the approaches of VC. In this test, the subjects were asked to choose between A and B based on the naturalness and similarity of the utterance compared to the target sample, X. The scores of this test are indicated as ABX, and naturalness in Figure 7.3 (a)-(b). On the similar lines, we also calculated the preference scores of these utterances using MCD and EE. Out of A and B that gave least value of MCD and EE, was preferred and the preference score (in %) are shown in Figure 7.3 (a)-(b). From Figure 7.3 (a)-(b), it can be seen that for both M-F and F-M VC, MCD gave 100% preference to GMM-based VC method. However, unlike MCD which gave 0% preference to BLFW+AS VC system, EE gave 45.8% preference score in case of F-M and 16.67% preference score in the case of M-F. Hence, EE is relatively more reliable than the MCD, which completely nullifies the possibility of BLFW+AS to be better in any case. Thus, we proposed the EE as an objective measures for assessing the quality of VC.

7.3.4 Chapter Summary

The study reported in this chapter investigated the novel objective measure which is based on the articulatory parameters. In particular, after VC, the articulatory parameters related information is lost which is quantified by a proposed objective measure, namely, EE. Though MCD and EE were found to be partially correlated and gave almost a similar kind of interpretation, EE had more correlation with MOS. The experiments showed that in the case of preference score, where MCD was 100 % contradicting subjective measure, which is highly unlikely. On the other hand, EE supported subjective measure 45.8 % and 16.67 % for F-M and M-F VC, respectively. Hence, the proposed measure EE is a reliable objective measure for measuring the quality of a voice converted speech. In the next chapter, mapping techniques of the VC have been applied to the different cross-domain conversion tasks.

CHAPTER 8

Cross-Domain Conversions

8.1 Introduction

In the last chapter, novel application of Acoustic-to-Articulatory inversion technique is proposed for the quality assessment of VC systems. In this chapter, mapping techniques of the VC techniques presented in Chapter 3 are extended for different cross-domain conversion applications of VC, namely, Non-Audible Murmur (NAM)-to-Whisper Speech, Whisper-to-Normal, and NAM-to-Normal Speech conversions. The murmur produced by the speaker and captured using the Non-Audible Murmur (NAM) microphone, one of the Silent Speech Interface (SSI) technique, suffers from degradation in speech quality and lacks in speech intelligibility. This is primarily due to the lack of radiation effect at the lips (which is approximated as highpass filtering), and the lowpass nature of the soft tissue, that attenuates the high frequency-related information. Hence, conversion of NAM signal to the whispered or natural speech proves to be a very challenging task [32].

In addition, NAM signal, whispered, and normal speeches are different modes of communication, from a speech production-perception viewpoint. In particular, during a normal speech production, air flowing through the lungs, causes the vocal folds to vibrate and as a result, this excitation source resonates the vocal tract [17]. However, during a whispered speech production, the glottis is open (i.e., no vocal fold vibrations), which causes the exhaled air to pass through the glottal constrictions [17]. This results into turbulent noisy excitation source for the vocal tract system [243]. On the other hand, signal that is recorded via NAM microphone, is the respiratory sound produced by the movement of articulators. NAM signal is recorded through the soft tissue of the head without any obstructions from the bones in the absence of vocal fold vibrations [244]. Furthermore, NAM signal, and whispered speech are completely aperiodic or unvoiced (due to turbulence created by geometry of narrow constriction [17]). In addition, the *non-linear* coupling between the vocal tract, and the excitation source increases in the whispered speech, that results in spectral differences between the normal and the whispered speech [17]. Hence, phone duration, energy distribution across phone classes, formant locations and the spectral tilt will also be affected, that alters the distribution of phones in the formant space, such as vowel triangle [245, 246]. Hence, NAM-to-SPeeCH (NAM2SPCH) conversion technique requires to find the mapping between such cross-domain entities.

The key goal of this work is to improve the intelligibility, and the naturalness of the NAM signal using a NAM-to-WHiSPer (NAM2WHSP) speech conversion technique. There are primarily two approaches for the NAM2WHSP conversion task. One based on speech recognition and synthesis [247] and the other that uses conversion or mapping techniques [248–250]. The first approach requires the linguistic information, which is unavailable for most of the cases. Moreover, our goal is to extract the linguistic message from the given silent speech. Hence, this work focuses on the second approach. It is natural to humans that they can easily recognize cross-domain relations, due to robust and sophisticated hearing mechanism [17]. However, it is difficult for machine to achieve a similar ability. The problem is to find mapping function between cross-domains, which is reformulated as generating a speech in one domain, given another speech from other-domain.

In this apply, we apply different deep learning-based mapping techniques (discussed in Chapter 3) of the VC to learn the cross-domain relations (w.r.t. attributes of the speech production mechanism) between NAM, and the whispered speech. Furthermore, we propose to use speaker-independent Whisper-to-Speech (WHSP2SPCH) conversion technique to further improve the naturalness of the converted whispered speech. There is an unavailability of normal speech corresponding to the whispered speech, and the NAM signal from the same speaker. Hence, we propose to develop a two-stage NAM-to-SPeeCH (NAM2SPCH) conversion model, where the first stage converts the NAM signal to the whispered speech, and the second stage converts the whispered speech to the normal speech.

Here, primary goal of the WHSP2SPCH system is to predict F_0 from the converted whispered speech. Attempts have been made in the past to predict F_0, directly or via statistical-based approaches for the WHSP2SPCH conversion tasks [24, 251–253]. It has been observed that the sensation of pitch (i.e., perceived pitch) still exists, which is encapsulated in an intricate way in the whispered speech [251, 254]. Hence, we predict F_0 from the cepstral features of the converted normal speech instead of predicting it directly from the cepstral features of the whispered speech (as suggested in [252]). Here, we apply proposed MMSE-

GAN, and the MMSE DiscoGAN architectures for the WHSP2SPCH task by analyzing the level of contextual information, and the number of training utterances required for optimizing the network parameters for the given task. The key contribution of this chapter is to apply proposed GAN-based architectures for the conversion tasks. Statistically meaningful analysis of subjective as well as objective evaluations are presented here for all the developed VC systems. In addition, F-ratio analysis has been used to illustrate how the proposed GAN-based frameworks incorporate the lip radiation-like smooth approximated *highpass* filtering effect via the conversion task [13,17].

8.2 Analysis of NAM Signal and Whispered Speech

In the absence of vocal fold vibrations, the respiratory sound produced by the movement of articulators, such as palate, tongue, lips, etc. can be recorded via the soft tissue of the head without any obstructions from the bones. This recorded signal is the non-audible murmur (NAM) [244].

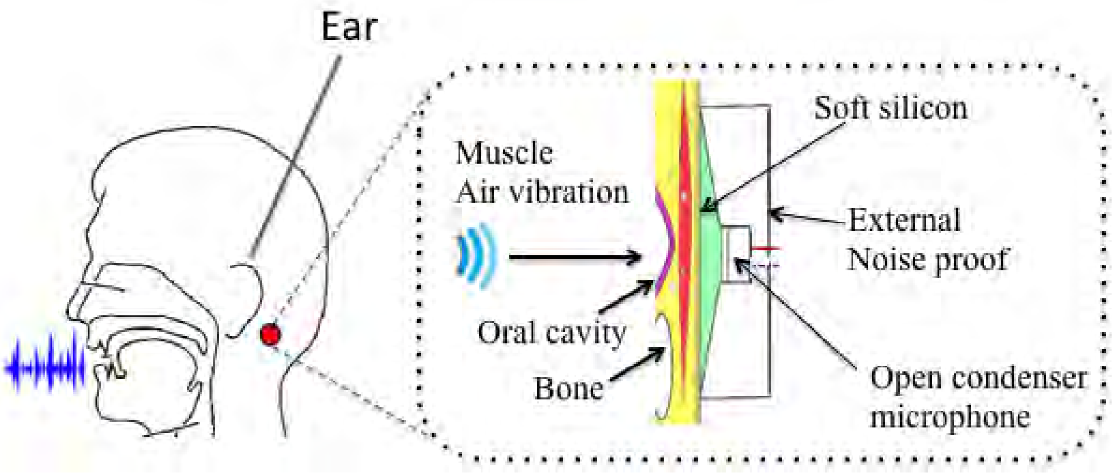


Figure 8.1: Schematic representation of NAM microphone. Adapted from [32,33].

The NAM is recorded by the special kind of stethoscopic microphone attached behind the ear to the surface of the skin (as shown in Figure 8.1). The NAM microphone contains an electret condenser, which is covered with urethane elastomer or soft silicon-like soft polymer material for better impedance matching with the soft tissue of the neck [32,255]. The acoustic vibration of the vocal tract system is captured via sensor placed close behind the ear, and it is then passed through the microphone. The Open Condenser Wrapped with Soft Silicon (OCWSS) type NAM microphone can record frequency upto 4 kHz [32].

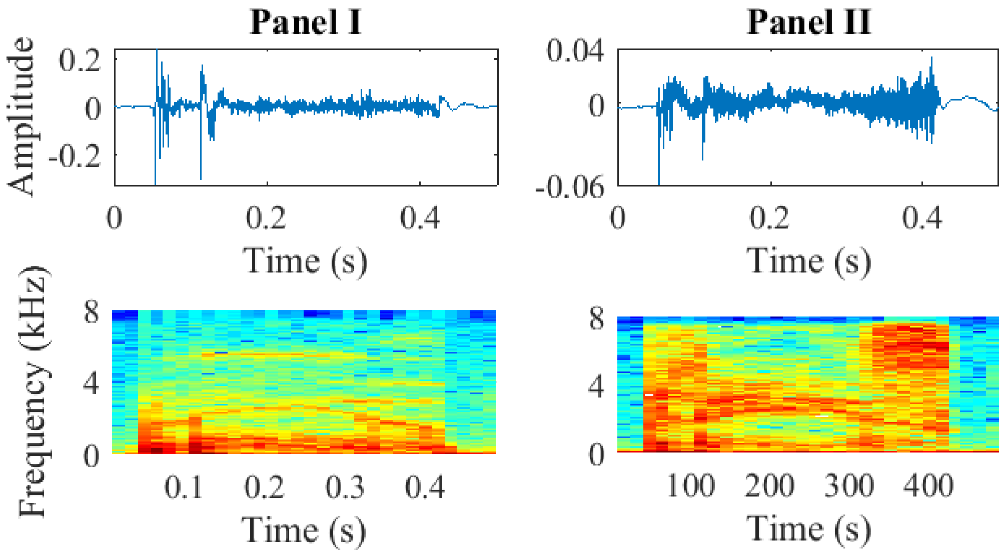


Figure 8.2: Time-domain signal, and its corresponding spectrogram for NAM signal (Panel I), and whispered speech (Panel II) for a word, /"Please"/. After [13].

On the other hand, turbulent noisy excitation source excites the vocal tract system, and produce the whispered speech in the absence of vocal fold vibrations [243]. Figure 8.2 shows the time-domain waveform, and its corresponding spectrogram for the NAM signal, and the whispered speech in Panel I, and Panel II, respectively. NAM signal is robust to the external noise compared to the normal speech or whispered speech, since it is recorded at the soft body tissue. Since body tissue act as a lowpass filter, high frequency information is attenuated in spectrogram (as shown in Panel I). On the other hand, the whispered speech is more sensitive to the external noise due to its recording at the lips. Lip radiation act as a highpass filter. Hence, high frequency information is present in the whispered speech spectrogram (as shown in Panel II).

According to the source-filter model, production of speech in discrete-time domain can be expressed as [17]:

$$S(z) = A_v G(z) V(z) R(z), \tag{8.1}$$

where $S(z)$ is the $\mathcal{Z}$-transform of speech signal, A_v is the gain which controls loudness, $G(z)$ is the $\mathcal{Z}$-transform of the glottal flow input, $V(z)$ is the vocal tract transfer function from the glottis to the lips, and $R(z)$ is the system function for the lip radiation effect. In normal speech, glottal source can be modelled as periodic, noisy, and impulsive depending on the speech sound (i.e., phone), speaker, and speaking style. However, in the case of the whispered speech, it will be always unvoiced (due to no vocal fold vibrations), and it can be represented as:

$$W(z) = A_w G_n(z) V(z) R(z), \tag{8.2}$$

where A_w is the gain corresponding to the loudness of the whispered speech, $W(z)$ is the $\mathcal{Z}$-transform of discrete-time whispered speech signal, and $G_n(z)$ is the noisy glottal source. Similarly, the unvoiced NAM signal from speech production can be represented as:

$$N(z) = A_n G_n(z) V(z) H(z), \tag{8.3}$$

where A_n is the gain corresponding to the loudness of the NAM signal, $N(z)$ is the $\mathcal{Z}$-transform of discrete-time NAM signal, and $H(z)$ is the impulse response of the NAM microphone, which is lowpass in nature. Here, since NAM is not recorded at the lips, the lip radiation effect will not be present. In particular, the lip radiation effect corresponds to transforming the volume velocity waveform at the lips into an acoustic pressure waveform (some distance away from the lips), which is caused by the boundary conditions of the vocal tract resonator tube [17, 256]. The lip radiation effect can be modeled as a first-order time-derivative Finite Impulse Response (FIR) filter (highpass filter) [17], i.e.,

$$R(z) = 1 - \alpha z^{-1}, \tag{8.4}$$

where α is a filter coefficient. Though $R(z)$ has a single zero on the unit circle in z-plane, it is modelled as a zero slightly inside unit circle for more realistic implementation [17]. Hence, α is taken between 0.98 to 0.99 [256]. The key goal of this work is to convert NAM signal into the whispered speech or normal speech via mapping techniques. Hence, the mapping techniques should incorporate the lip radiation-like smooth highpass filtering effect, which is absent in the NAM signal.

8.3 Proposed NAM-to-WHiSPer (NAM2WHSP) Conversion

The block diagram of the NAM2WHSP conversion system is shown in Figure 8.3. First, the cepstral features are extracted from the NAM signal, and its corresponding whispered speech. In this work, we have applied three conversion techniques (as summarized in Section II) to learn the mapping function between the cepstral features of the NAM signal, and the whispered speech. As both the NAM signal, and the whispered speech are unvoiced sounds, we did not apply F_0 conversion techniques. At the time of testing, we extracted the cepstral features from the input NAM signal, and converted it into cepstral features of the whispered speech using the learned mapping function. At the end, features are converted back into the whispered speech using the VOCODER.

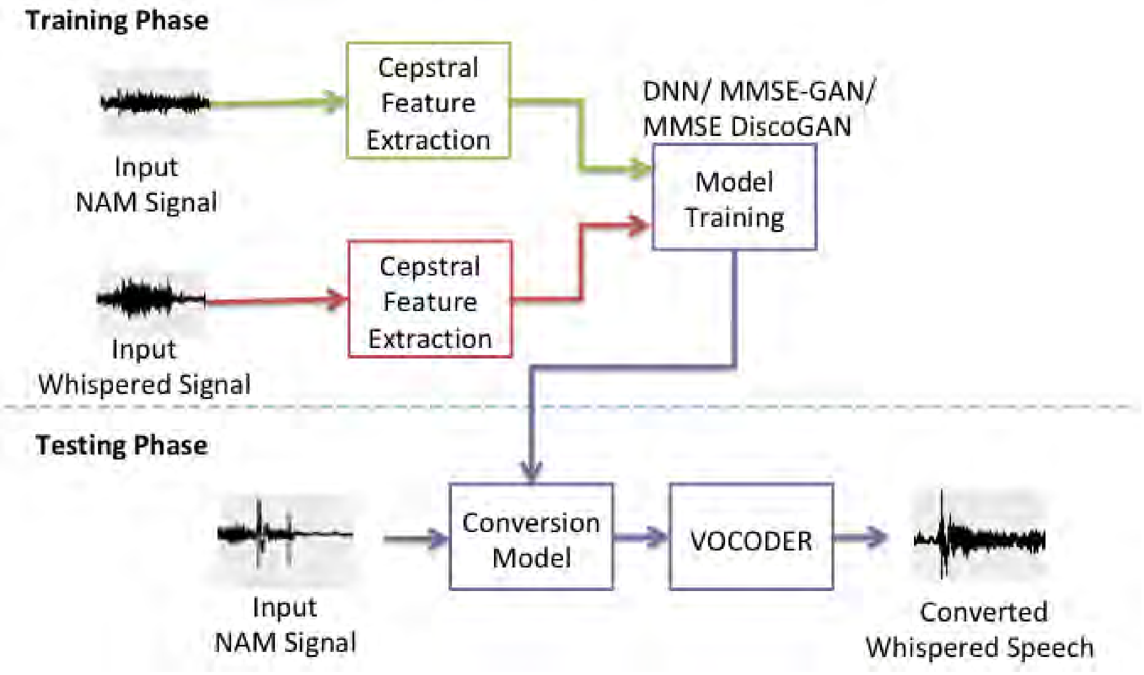


Figure 8.3: Schematic representation of the NAM2WHSP conversion system. After [13,34].

8.4 Proposed WHiSPer-to-SPeeCH (WHSP2SPCH) Conversion

Due to unavailability of normal speech corresponding to the NAM, and the whispered speech from the same speaker, we propose to develop *speaker-independent* WHSP2SPCH conversion system to further improve the naturalness of the converted whispered speech. We use multiple speakers' whispered speech, and its corresponding normal speech to develop speaker independent WHSP2SPCH conversion system. Conversion techniques for cepstral features are almost the same as mentioned in the NAM2WHSP conversion (in Section 8.3). However, one of the main issue before learning the mapping function for the WHSP2SPCH conversion system is the time-alignment between whispered, and its corresponding normal speech. To that effect, we have used Dynamic Time Warping (DTW) algorithm [248,250].

Conversion techniques are applied to learn 1) mapping between the cepstral features corresponding to the whispered and normal speeches, and 2) the mapping between the cepstral features, and the corresponding F_0 of the normal speech. At the time of testing, we predicted F_0 from the cepstral features of the converted normal speech instead of predicting directly from the cepstral features of the whispered speech (as suggested in [252]). In addition, we have used Artificial Neural Network (ANN) for deciding voiced-unvoiced frames. At the end, these converted cepstral features, and the predicted F_0 are converted to the normal speech using VOCODER.

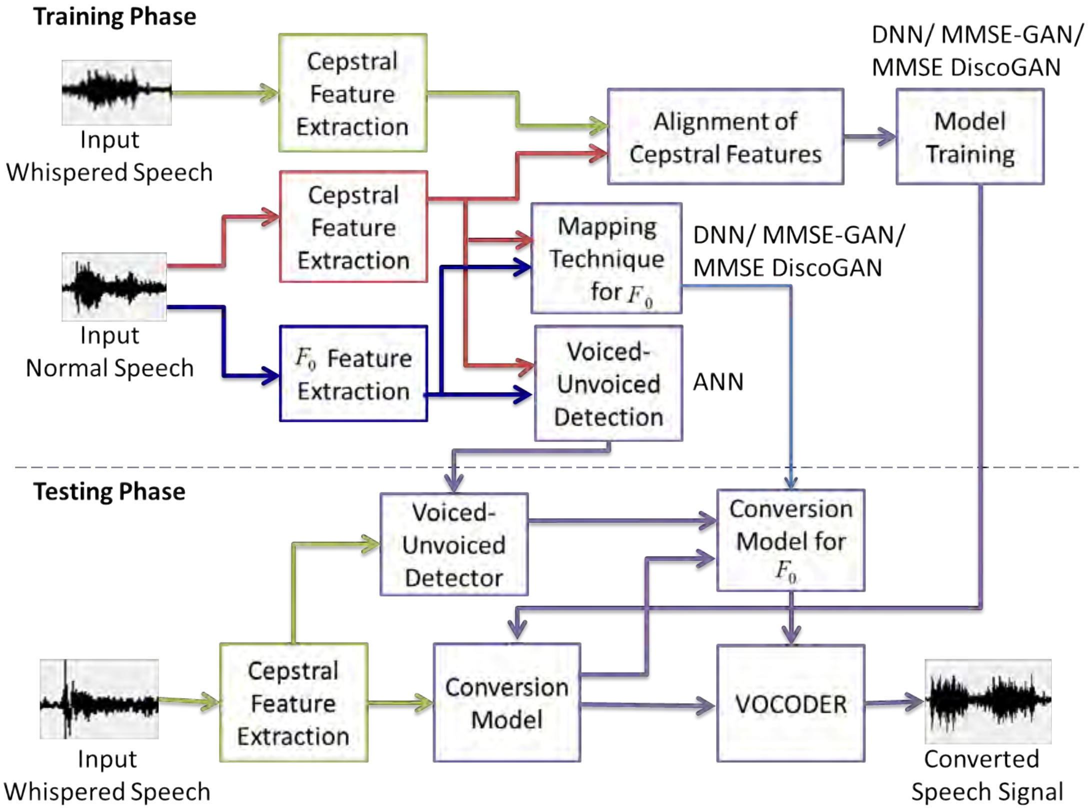

Training Phase
Input Whispered Speech
Cepstral Feature Extraction
Alignment of Cepstral Features
DNN/ MMSE-GAN/ MMSE DiscoGAN
Model Training
Input Normal Speech
Cepstral Feature Extraction
F_0 Feature Extraction
Mapping Technique for F_0
DNN/ MMSE-GAN/ MMSE DiscoGAN
Voiced-Unvoiced Detection
ANN
Testing Phase
Voiced-Unvoiced Detector
Conversion Model for F_0
Input Whispered Speech
Cepstral Feature Extraction
Conversion Model
VOCODER
Converted Speech Signal

Figure 8.4: Schematic representation of the WHSP2SPCH conversion system. After [13,24].

8.5 Proposed NAM-to-SPeeCH (NAM2SPCH) Conversion

Once both the NAM2WHSP and the WHSP2SPCH systems are trained, we combine both the systems to develop the proposed NAM2SPCH conversion system. The basic block diagram of the proposed NAM2SPCH conversion system is shown in Figure 8.5. Here, from a given NAM signal, first we extract cepstral features corresponding to the NAM signal, and then convert it into cepstral features corresponding to the whispered speech using the NAM2WHSP system. These converted cepstral features are then passed through Voiced-Unvoiced Detector (VUD). On the other hand, the converted cepstral features obtained after the NAM2WHSP systems are further passed through the speaker-independent WHSP2SPCH conversion system, and these cepstral features are then converted into cepstral features of the normal speech signal. F_0 contour is then predicted using the converted cepstral features obtained from the WHSP2SPCH system for the voiced frames. At the end, both the cepstral features, and its predicted F_0 values are passed through the VOCODER, and converted into the audible normal speech signal.

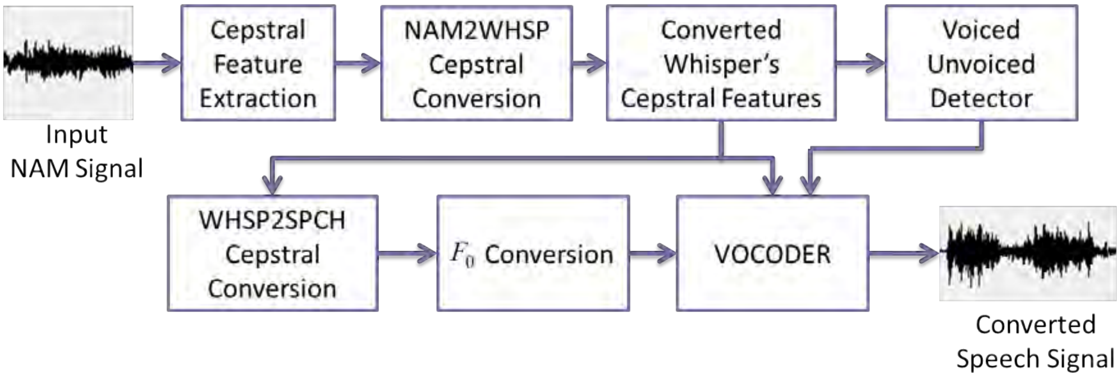

Input
NAM Signal
Cepstral Feature Extraction
NAM2WHSP Cepstral Conversion
Converted Whisper's Cepstral Features
Voiced Unvoiced Detector
WHSP2SPCH Cepstral Conversion
F_0 Conversion
VOCODER
Converted
Speech Signal

Figure 8.5: Schematic representation of the proposed NAM2SPCH conversion system. After [13].

8.6 Experimental Results

This section will present the details of the experimental setup and the results obtained for the above mentioned three conversion techniques.

8.6.1 Experimental Setup

For the development of NAM2WSHP conversion systems, 420 NAM signals, and their corresponding whispered speech utterances of Herald Glasgow newspaper texts have been selected from the CSTR NAM TIMIT Plus corpus [257]. Out of these 420 utterances, we have selected 400 utterances for the training and the remaining 20 utterances for the testing. In order to develop speaker-independent WHSP2SPCH conversion system, we have used in total 40 speakers' whispered speech, and its corresponding normal speech data from the two databases, namely, The CHAracterizing INdividual Speakers (CHAINS) Speech Corpus [258], and the Electromyographic (EMG)-UKA Trail corpus [259]. We have taken in total 1302 utterances for the training, and 108 utterances for the testing.

In this thesis, the DNN, the G network in GAN, and the generators (i.e., G_{AB} and G_{BA}) of MMSE DiscoGAN follows the identical architecture, with the three hidden layers. Having a uniform architecture helps in analyzing the advantages of adversarial training over the MMSE-based ML optimization. Each hidden layer contains 512 neurons with Rectified Linear Unit (ReLU) activation, whereas the output layer has the linear activation function. The D network in the MMSE-GAN and the discriminators (i.e., D_A and D_B) of the MMSE DiscoGAN also have three hidden layers, with *tanh* activation function, whereas the output layer has sigmoid activation as suggested in [24,34,48]. All the three models are trained for 100 epochs, using an effective batch size of 1000 frames as suggested in [34, 48]. The parameters are optimized using an Adam optimization [98], with a learning rate of 0.001. The *25*-dimensional Mel Cepstral Coefficients (MCCs) (including the 0^{th} coefficient) are extracted from the NAM, whispered, and normal speeches with 25 ms Hamming window, and 5 ms overlap between the consecutive frames. For analysis-synthesis, we have used AHOCODER [67].

8.6.2 Objective Evaluation

We have applied appropriate measures-based on the objectives of the conversion system. In particular, we have applied Mel Cepstral Distortion (MCD), and Perceptual Evaluation of Speech Quality (PESQ) to measure the effectiveness of the NAM2WHSP conversion systems. The traditional MCD measure is used here which is given by [16]:

$$MCD \text{ [in dB]} = \frac{10}{ln10}\sqrt{2\sum_{i=1}^{25}(m_i^r - m_i^c)^2} \ , \tag{8.5}$$

where m_i^r and m_i^c are the i^{th} MCCs of the reference, and the converted signal. In particular, m_i^r and m_i^c are the i^{th} MCCs of the reference whispered speech, and the converted whispered speech in the case of NAM2WHSP conversion system, and the reference normal speech, and the converted normal speech in the case of WHSP2SPCH conversion system. Since MCD is the distance between the converted and the reference cepstral features, a system that is having lesser MCD is considered as relatively better system.

The PESQ score is a prominent objective measure for evaluating quality of speech objectively [260]. It is computed by taking the linear combination of d_{sym} (average normal disturbance value), and d_{asym} (average assymetrical disturbance value) between the reference and the converted signal, which is given by [260]:

$$PESQ = 4.5 - 0.1 \times d_{sym} - 0.0309 \times d_{asym}. \tag{8.6}$$

A system that is having higher PESQ score is considered as relatively better system. Since the converted whispered speech, and the reference whispered speech are time-aligned, we measure the PESQ only for the evaluations of NAM2WHSP conversion system. However, the converted normal speech obtained using the WHSP2SPCH conversion system, and the reference normal speech are not time-aligned. Thus, we did not use the PESQ as an objective measure for this system. Since the primary goal is to accurately predict fundamental frequency (i.e., F_0) from the unvoiced whispered speech [251], we have primarily considered Root Mean Square Error (RMSE) of $\log(F_0)$ as an objective measure for the WHSP2SPCH conversion systems. To measure the RMSE of $\log(F_0)$, the converted speech signal, and the actual reference speech signal are time-aligned using the DTW algorithm. These DTW aligned pairs will generate voiced-voiced, voiced-unvoiced, unvoiced-voiced, and unvoiced-unvoiced pairs. Here, we consider only voiced-voiced pairs for computing the RMSE of the $\log(F_0)$ (since F_0 is undefined for the unvoiced frames) [114]. RMSE of the $\log(F_0)$ is given by:

$$RMSE(\log(F_0)) = \sqrt{\sum_{i=1}^{k} [\log(F^r_{0_i}) - \log(F^c_{0_i})]^2}, \tag{8.7}$$

where k is the total number of voiced-voiced pairs after the alignment, and ${F_0}^r$ and ${F_0}^c$ are the F_0 of the reference and the converted speech signals, respectively. Lesser the RMSE of $\log(F_0)$, better the system is.

8.6.2.1 Effect of Contextual Information

Extracting the contextual features from the speech are necessary due to its sequential nature. In addition, contextual features capture the local features (including coarticulation), and preserve the crucial harmonics, [261, 262]. In speech perception studies, it has been shown that the surrounding acoustic context impacts the human perception [263–265]. Recently, studies in neuroscience have tried to identify the underlying representations in the primary, and secondary auditory cortex, and examined the information modulated by varying the context [199]. Motivated from this work, we analyze the effect of varying number of contextual frames (i.e., 0, 1, 2, 3, 4, and 5 that will correspond to the window length 1, 3, 5, 7, 9, and 11 frames) at the network input in a symmetric way. We also analyze the importance of training models by taking an asymmetric contextual frames as an input to the network, as suggested in [262]. These symmetrical and asymmetrical contextual NAM speech parameters will predict the 25-dimensional MCCs of the whispered speech via NAM2WHSP system.

Figure 8.6 (a) and (b) shows the MCD, and the PESQ score of the NAM2WHSP conversion system along with 95 % confidence interval for different symmetric contexts. Effectiveness of the proposed MMSE-GAN, and MMSE DiscoGAN-based architectures over the baseline DNN-based architectures can be clearly seen in both the MCD as well as the PESQ scores. In particular, the proposed MMSE-GAN, and MMSE DiscoGAN obtained on an average 6.30 % and 6.16 % relative reduction in the MCD, and 6.15 % and 5.77 % relative improvement in the PESQ w.r.t. the baseline DNN-based systems, respectively. It can be clearly seen from Figure 8.6 (a) and (b) that the performance of the proposed systems for the symmetric context with 4 frames is on an average relatively better in terms of both the PESQ and the MCD.

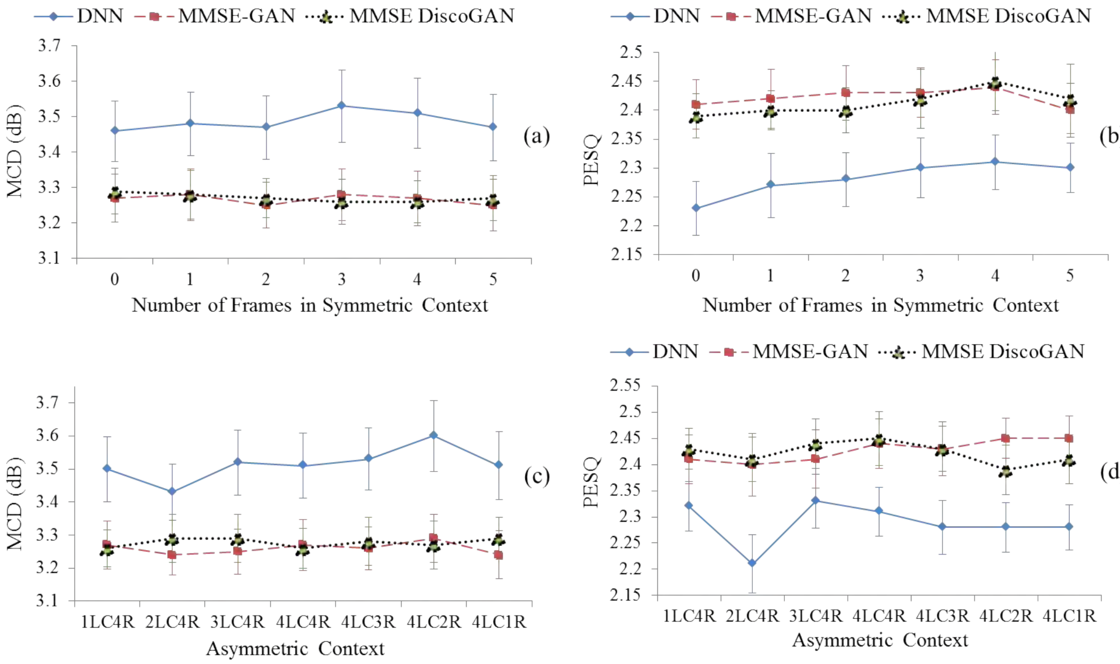


Figure 8.6: MCD and PESQ analysis of different NAM2WHSP systems for (a), (b) symmetric, and (c), (d) asymmetric contextual input frames. After [13].

Figure 8.6 (c) and (d) show the MCD, and PESQ of the NAM2WHSP conversion system for the different asymmetric contexts. There is on an average 7.22 % and 6.73 % relative reduction in the MCD, and 6.14 % and 5.95 % relative improvement in the PESQ with the MMSE-GAN, and the MMSE DiscoGAN-based architectures over the baseline, respectively. Furthermore, the relative improvement of the proposed architectures over the baseline can be observed from Figure 8.6 for the cases, where we took more number of frames in the left context (i.e., asymmetric left context for the systems 4LC3R, 4LC2R, 4LC1R) than that for the right context (i.e., asymmetric right context for the systems 1LC4R, 2LC4R, 3LC4R) in the asymmetric combinations. In particular, relative reduction in the MCD, and the relative improvement in the PESQ for the proposed GAN-based architectures over the baseline DNN are shown in Table 8.1 and Table 8.2, respectively.

Table 8.1: % Relative reduction in the MCD for the proposed GAN architectures over the baseline DNN. After [13]

	Asymmetric Right	Symmetric	Asymmetric Left
MMSE-GAN	6.59 %	6.83 %	**7.98** %
MMSE DiscoGAN	5.82 %	7.12 %	**7.51** %

Table 8.2: % Relative improvement in the PESQ for the proposed GAN architectures over the baseline DNN. After [13]

	Asymmetric Right	Symmetric	Asymmetric Left
MMSE-GAN	5.30 %	5.63 %	**7.16** %
MMSE DiscoGAN	**6.17** %	6.06 %	5.70 %

Table 8.3: % Relative reduction in the RMSE of $\log(F_0)$ for the proposed GAN architectures over the baseline DNN. After [13]

	Asymmetric Right	Symmetric	Asymmetric Left
MMSE-GAN	31.46 %	27.78 %	**33.95** %
MMSE DiscoGAN	31.18 %	32.04 %	**37.83** %

For the WHSP2SPCH conversion systems, significant reduction in the RMSE of $\log(F_0)$ is observed for the symmetric as well as the asymmetric contexts (as shown in Figure 8.7 (a) and Figure 8.7 (b)) indicating effectiveness of the proposed GAN architectures. In particular, we obtain on an average 26.17 % and 31.99 % relative reduction in the RMSE of $\log(F_0)$ over the baseline DNN in the symmetric context for MMSE-GAN and the MMSE DiscoGAN, respectively. In addition, % relative reduction in the RMSE of $\log(F_0)$ increases with the number of frames in the symmetric context. Similarly, there is on an average 32 %, and 34.15 % relative reduction in the RMSE of $\log(F_0)$ over the DNN for different cases of asymmetrical context with the MMSE-GAN, and the MMSE DiscoGAN, respectively. Furthermore, relative reduction in the RMSE of $\log(F_0)$ for the proposed GAN-based architectures over the baseline is shown in Table 8.3.

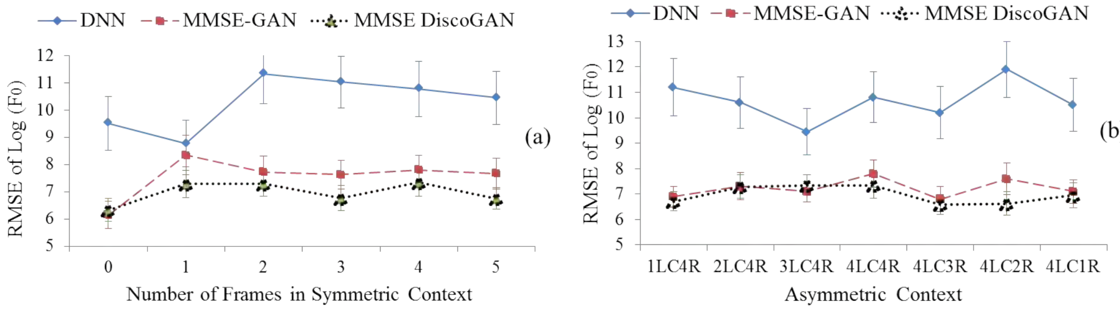


Figure 8.7: RMSE of $\log(F_0)$ for the different WHSP2SPCH systems for (a) symmetric, and (b) asymmetric contextual input frames. After [13].

8.6.2.2 Analysis w.r.t. Amount of Training Data

We analyze the models for varying size of the training data from the 10000 frames (less amount of training data) to the 40.6 lacs frames (relatively a large amount of training data). Figure 8.8 (a), Figure 8.8 (b), and Figure 8.8 (c) show that the MMSE-GAN, and the MMSE DiscoGAN achieves better MCD, PESQ, and RMSE of the $\log(F_0)$, as the number of training utterances increases. In particular, MMSE-GAN, and MMSE DiscoGAN achieve on an average -4.47 % to the 6.84 %, and -5.58 % to 7.12 % relative reduction in the MCD values over the baseline DNN, respectively. Similarly, observations have been perceived in PESQ score as well. Furthermore, in the case of WHSP2SPCH conversion systems, we obtain on an average 17.15 % to 26.60 %, and 10.49 % to the 35.40 % relative reduction in the RMSE of $\log(F_0)$ for the MMSE-GAN, and MMSE DiscoGAN w.r.t. baseline DNN for different amount of training data, respectively.

Initially, the DNN had ascendancy over the GAN-based architectures. However, the objective scores predicted by the DNN deteriorates significantly as the amount of training data increases. This may be due to the fact that, with the less amount of training data, the discriminators in the MMSE-GAN, and the MMSE DiscoGAN are very confident about its decision of rejecting the generated cepstrum. Moreover, with generator exposed to a very few amount of training data (though indirectly), it may not be able to sufficiently *fool* the discriminator. However, with the exposure to the large amount of training data, the adversarial training forces the generator to produce the cepstrum that approximately follows the data distribution and thus, successfully confuses the discriminator.

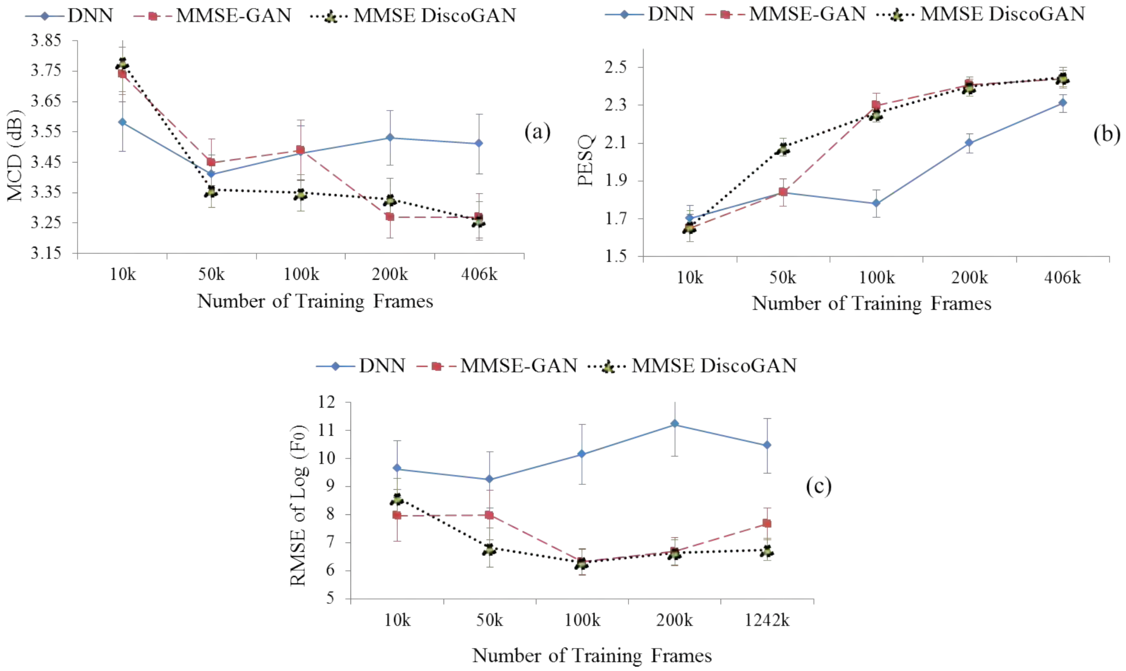


Figure 8.8: (a) MCD, (b) PESQ, and (c) RMSE of $\log(F_0)$ analysis of the various developed NAM2WHSP, and WHSP2SPCH systems w.r.t. the amount of available training data. 'k' in X-axis labels of each sub-figure corresponds to multiplier of 1000. After [13].

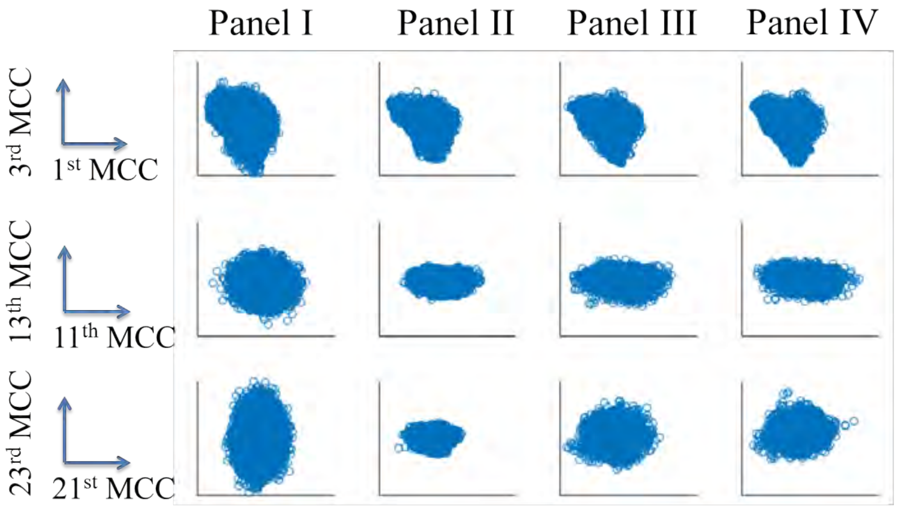


Figure 8.9: Scatter plots of several randomly selected pairs of the MCC's dimensions for NAM2WHSP systems. Panel I: Whispered speech, Panel II: DNN, Panel III: Proposed MMSE-GAN, and Panel IV: Proposed MMSE DiscoGAN. After [13].

8.6.2.3 Distributions of the Generated Parameters

To further investigate the effectiveness of the proposed GAN-based NAM2WHSP systems over the baseline DNN-based NAM2WHSP systems, we present the scatter plots (as shown in Figure 8.9) for the several MCCs pairs obtained from the converted whispered speech, and the ground truth (i.e., the whispered speech) (as done in [145]). It can be observed that the converted parameters obtained from the DNN-based NAM2WHSP systems are narrowly distributed (i.e., having less scatter) for the MCCs corresponding to the higher frequency regions as compared to the lower frequency counterparts. However, the converted parameters obtained from the proposed MMSE-GAN, and MMSE DiscoGAN-based NAM2WHSP systems are widely distributed, and very close to the distribution of parameters corresponding to the ground truth.

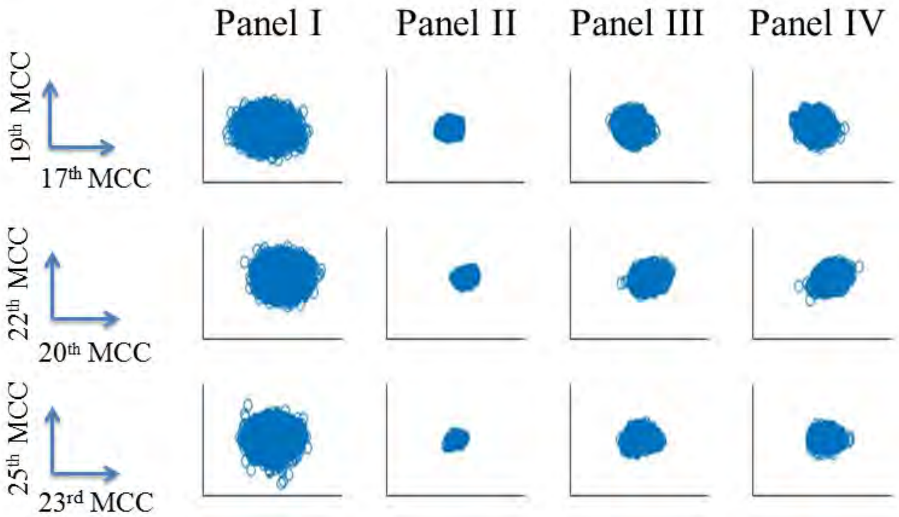


Figure 8.10: Scatter plots of pairs of the MCC's corresponding to the high frequency regions for NAM2WHSP systems. Panel I: Whispered speech, Panel II: DNN, Panel III: Proposed MMSE-GAN, and Panel IV: Proposed MMSE DiscoGAN. After [13].

In particular, Figure 8.10 shows the effectiveness of GAN-based architectures in generating the MCCs corresponding to the higher frequency regions over the baseline DNN. Similarly, the distribution of the predicted $\log(F_0)$ is presented in Figure 8.11 for evaluating the performance of the WHSP2SPCH conversion systems. We can clearly see from the Figure 8.11 that the predicted $\log(F_0)$ from the MMSE-GAN, and the MMSE DiscoGAN clearly follows the distribution of the $\log(F_0)$ corresponding to the normal speech compared to the DNN-based WHSP2SPCH systems, indicating effectiveness of the proposed GAN architectures.

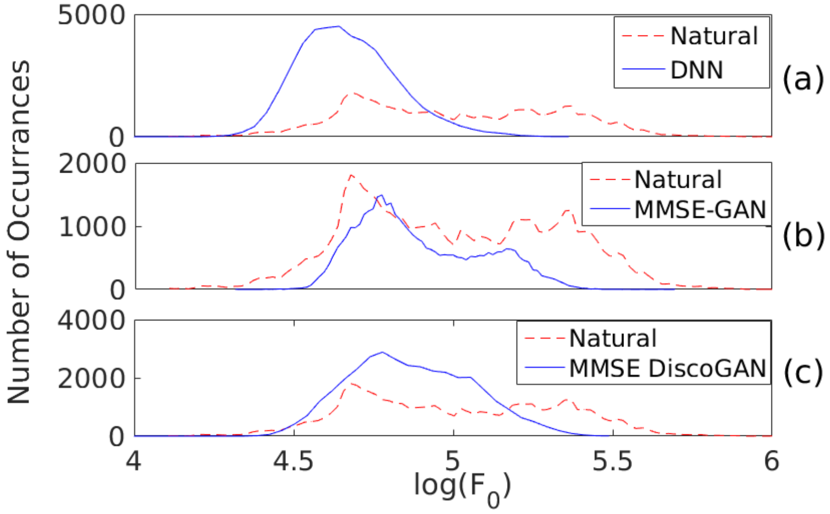


Figure 8.11: Histogram of generated $\log(F_0)$ for (a) DNN-based, (b) proposed MMSE-GAN-based, and (c) proposed MMSE DiscoGAN-based WHSP2SPCH systems along with the histogram of $\log(F_0)$ of natural speech signal. After [13].

8.6.3 F-ratio Analysis

The Fisher's F-ratio can be used to determine which frequency band is relatively more discriminative to separate the two classes. It is a ratio of the inter-class variance to the intra-class variance [162]. Hence, higher F-ratio means the better separation of the two classes. Here, we used F-ratio-based analysis to understand the effect of mapping function on the input NAM signal. In particular, we computed F-ratio between a class corresponding to the NAM signal (C_n), and a class corresponding to the converted whispered voices (C_c) across all the available frequency regions. Here, we used the cepstral features instead of the spectral features to calculate the F-ratio, which is given by [167]:

$$F_i = \frac{(\mu_i^n - \mu_i^c)^2}{\frac{1}{N_n}\sum_{x_i \in C_n}(x_i - \mu_i^n)^2 + \frac{1}{N_c}\sum_{x_i \in C_c}(x_i - \mu_i^c)^2}, \tag{8.8}$$

where x_i is the i^{th} cepstral coefficient of the MCC feature vector, $\mathbf{X} = [x_1, x_2, ..., x_d]$. In addition, N_n and N_c are the total number of frames in the class C_n and C_c, respectively. The μ_i^n and μ_i^c are the mean of the x_i of all the frames corresponding to the NAM signal, and the converted speech, respectively. The F-ratio for different frequency regions will form the F-ratio pattern, i.e., $[F_1, F_2, ..., F_d]$, where d is the dimension of the MCC feature vector.

Figure 8.12 shows the F-ratio analysis for the DNN, MMSE-GAN, and MMSE DiscoGAN-based mapping function. We can clearly see that F-ratio is higher for the MCCs corresponding to the lower frequency regions compared to the higher frequency counterparts. This clearly indicates that the converted cepstral features obtained after the mapping function have relatively more discrimination for the MCCs corresponding to the lower frequency regions than their high frequency counterpart. This may be due to the fact that our mapping function is significantly attenuating the low frequency region. This could possibly mean that our learned mapping function is indeed playing role of approximating the lip radiation effect in the form of highpass filtering [17].

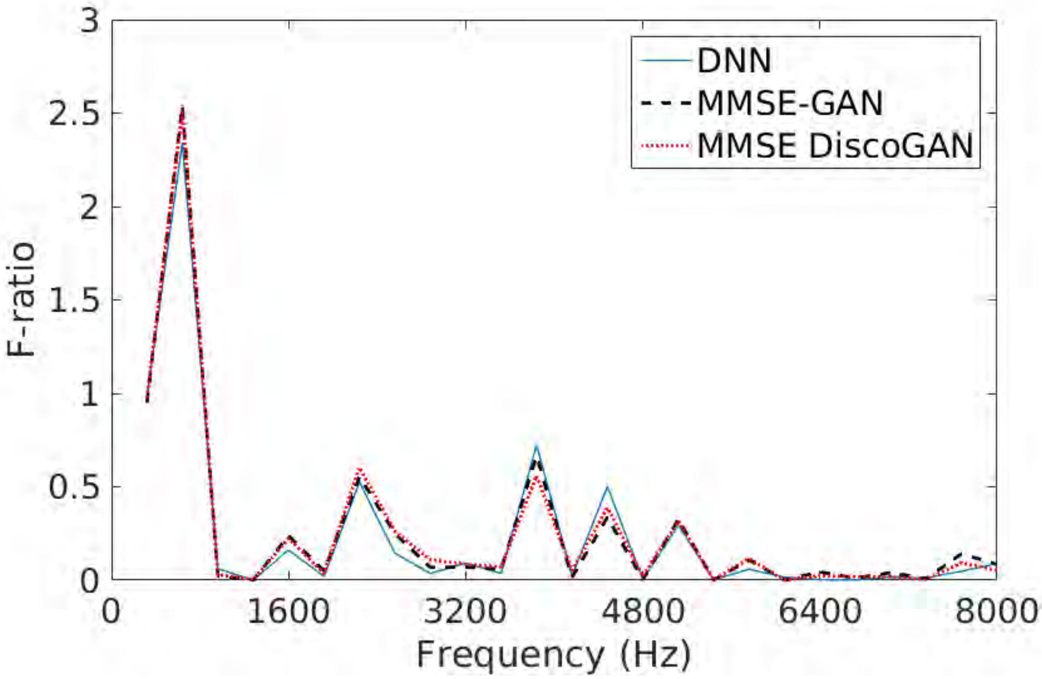


Figure 8.12: F-ratio pattern between the class corresponding to the original NAM signal, and the converted whispered speech. After [13].

8.6.4 Subjective Evaluation

Due to unavailability of natural speech corresponding to the NAM signal, we could not apply any objective measure for the NAM2SPCH conversion system. In addition, the key aim of the proposed work is to extract the linguistic content from the NAM signal, and improve its naturalness via NAM2WHSP, and NAM2SPCH (i.e., combined NAM2WHSP and WHSP2SPCH) conversion systems. Hence, we primarily focused on the subjective evaluations for our final NAM2WHSP and NAM2SPCH conversion system. For the subjective evaluations, we consider Mean Opinion Scores (MOS) test for intelligibility, and ABX test for naturalness. Total 40 listeners (14 females and 26 males without any known hearing impairments, and with age ranging from 18 to 29 years) took part in all the subjective tests. We used high quality Sennheiser headphones for the subjective evaluations. In MOS intelligibility test, subjects were asked to rate the played utterance from the various systems on the scale of 1−5, where 1 means not at all intelligible to 5 means completely intelligible. The results of the MOS test along with the 95 % confidence intervals are shown in Figure 8.13. We can clearly see from Figure 8.13 that the intelligibility of NAM signal is very less and hence, it is hardly intelligible. In both NAM2WHSP and NAM2SPCH cases, we can clearly see the effectiveness of the proposed GAN-based architectures over the baseline DNN. Relative improvements of the proposed architectures over the baseline in terms of MOS scores for the intelligibility are shown in Table 8.4.

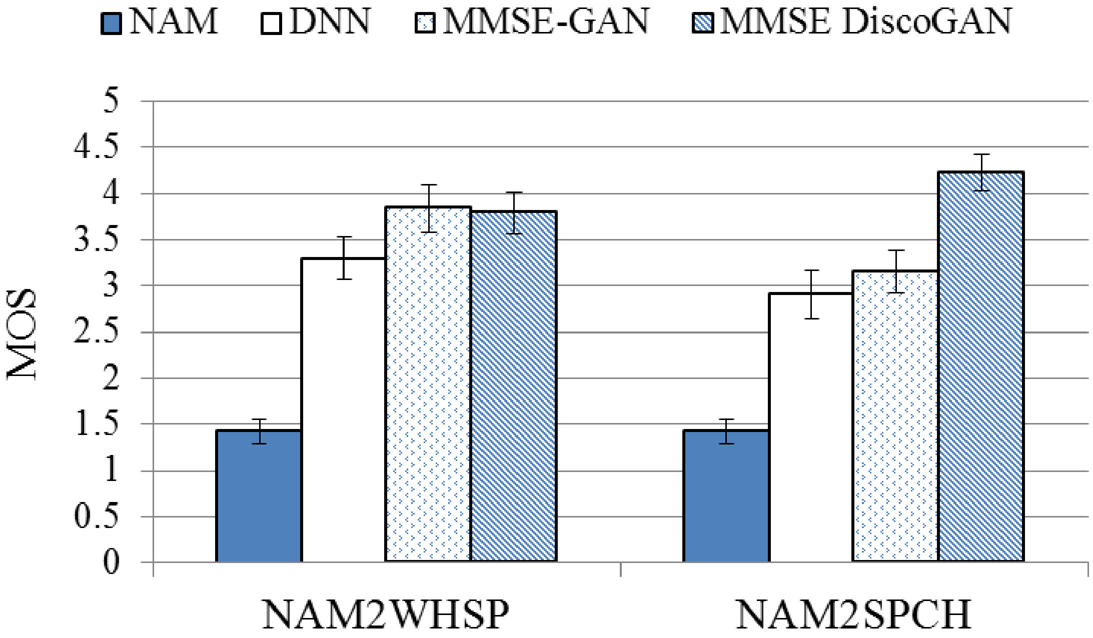


Figure 8.13: MOS scores for the intelligibility for both NAM2WHSP, and NAM2SPCH conversion systems. After [13].

Table 8.4: % Relative improvements in the MOS for the proposed architectures over the baseline DNN. After [13]

	NAM2WHSP	NAM2SPCH
MMSE-GAN	16.36 %	8.25 %
MMSE DiscoGAN	14.85 %	45.36 %

In ABX test, we randomly played the same utterances from two different systems, and asked subjects to decide which one is more better in terms of naturalness. Results of the ABX test for both the NAM2WHSP and NAM2SPCH are shown in Figure 8.14. We obtained on an average 0.11 margin of error corresponding to the 95 % confidence interval in the ABX test. It is clearly observed that in both the cases, proposed GAN-based systems are relatively more preferred by the subjects compared to the baseline DNN-based systems.

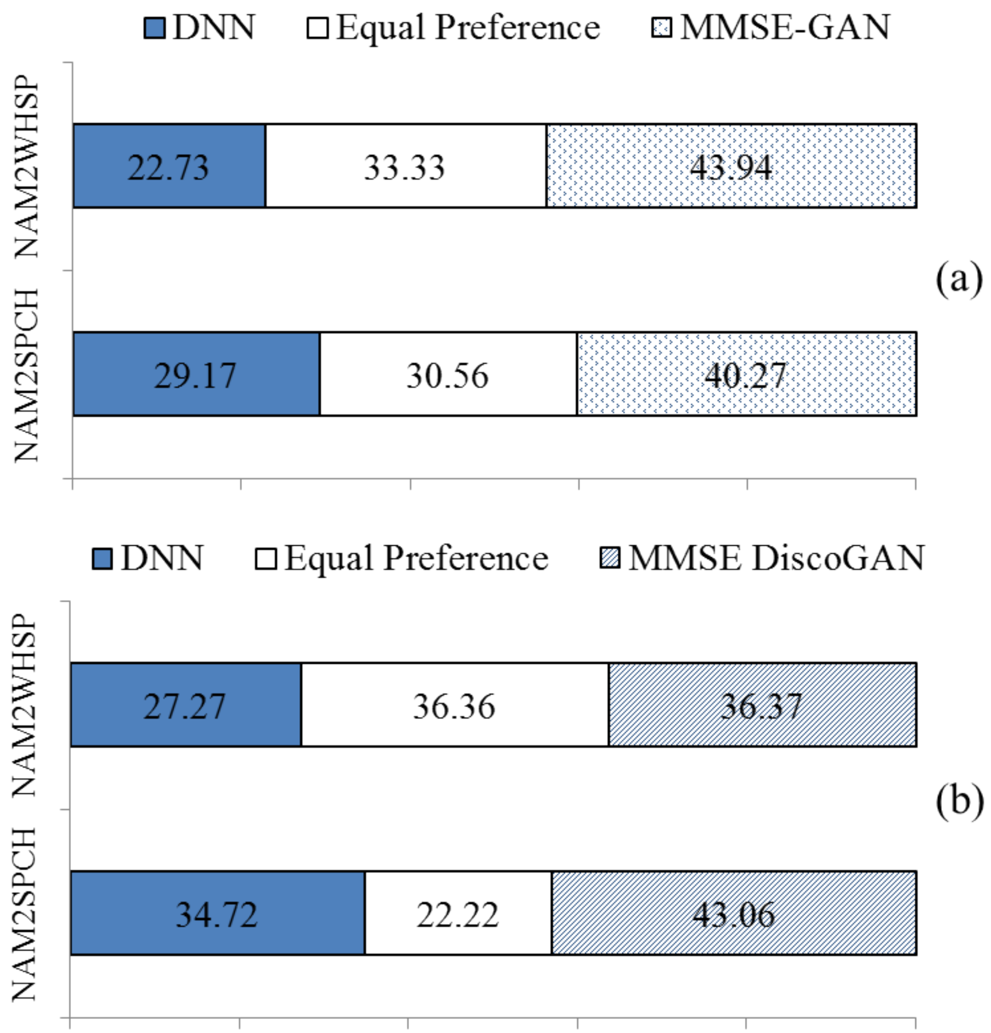


Figure 8.14: ABX test for naturalness for: (a) DNN *vs.* MMSE-GAN, and (b) DNN *vs.* MMSE DiscoGAN along with on an average 0.11 margin of error corresponding to the 95 % confidence interval. After [13].

It should be noted that the poor performance measure exhibited by the DNN (in both objective as well as subjective evaluations), may be primarily due to the absence of an *adversarial* nature of training (as in ML-based optimization). The DNN predicts the estimated output and with the true labels exposed, the network numerically optimizes the parameters, and then the network would be exposed to the unseen examples. However, the reduction in the numerical estimates does not always correlate with the generated sample quality [41]. Moreover, the ML-based optimization of the network parameters does not always lead to the better perceptual speech quality [266]. The ML-based optimization criteria put prior assumptions on the data distribution (such as, the MMSE objective function assumes that the output variables to be Gaussian), which may not be valid for the given data. Hence, such assumptions prevent the network to learn perceptually optimal network parameters for several speech technology applications [48]. On the other hand, GANs minimize the distributional divergence in learning the mapping function with adversarial training, and preserve the speech quality and achieve better intelligibility over DNN-based architectures.

8.7 Chapter Summary

In this chapter, we proposed the MMSE-GAN and the MMSE DiscoGAN architectures for finding mapping between cross-domain (w.r.t. the speech production) entities, namely, NAM and the whisper, via the NAM2WHSP task. In addition, we proposed to use the speaker-independent WHSP2SPCH conversion technique to further improve the naturalness of the converted whispered speech. In particular, our proposed NAM2SPCH conversion model consists of the two stages, namely, the NAM2WHSP and the WHSP2SPCH, due to unavailability of the normal speech corresponding to the NAM signal and the whispered speech from the same speaker. The proposed GAN architectures obtained better performance compared to the baseline DNN in both the objective as well as the subjective evaluations. This clearly indicates the effectiveness of GANs in minimizing the distributional divergence in learning the mapping function and preserving the speech quality and achieving better intelligibility gains over the traditional ML-based optimization techniques.

In addition, we identify the impact of symmetric as well as asymmetric contextual frames, and the number of training utterances required for optimizing the network parameters. We observed that in the case of less amount of training data, the DNN-based systems were performing better than the GAN-based systems. It is possibly due to the fact that, with the few training utterances, the discriminators in the MMSE-GAN and the MMSE DiscoGAN are confident about its decision of rejecting the generated cepstral features. However, the objective scores predicted by the DNN deteriorates significantly as the number of training utterances increases. This may be due to the fact that with the increased exposure of training utterances, the adversarial training forces the generator to produce the cepstrum that approximately follows the data distribution, and successfully confuses the discriminator. In addition, the effectiveness of the mapping function to incorporate the lip radiation-like smooth highpass filtering effect to improve the performance of proposed NAM2SPCH conversion system has shown via the F-ratio analysis. In the future, we plan to further improve the quality of the converted voices by employing high quality wavenet vocoder. The perceptual difference observed between the estimated and the ground truth indicates the need of exploring the better objective function that can perceptually optimize the network parameters. In next chapter, overall summary of the thesis work is presented.

CHAPTER 9

Summary and Conclusions

In this chapter, the overall summary of the thesis work is presented along with the limitations of the current research work, and potential future research directions.

9.1 Summary of Thesis Work

Following is the summary of the research work carried out in the entire thesis.

- In this thesis, the introduction, and motivation for the problem of VC are presented in Chapter 1 followed by a discussion on its various potential applications along with its basic system architectures. Background studies needed to understand the different stages of the VC technique is presented in Chapter 2. In addition, Chapter 2 also presents brief history and selected chronological progress in the VC field from different perspectives. Major contributions from mapping and alignment perspectives of the VC problem are presented in this thesis work.

- Contributions of the thesis from the mapping perspectives, and the state-of-the-art mapping techniques are presented in Chapter 3. In particular, the novel AS technique is presented for the BLFW-based VC, which exploits the advantage of GMM-based VC to obtain better speech quality compared to the state-of-the-art AS technique. Furthermore, DNN-based techniques are discussed in this thesis, and to tackle the issues related to overfitting, proposed modifications to the DNN architectures is presented. In particular, an empirical analysis is presented to show that with the proposed DNN architecture, the need for pretraining can be avoided in the context of VC without compromising speech quality, and speaker similarity. Finally, the proposed MMSE regularized GAN, and DiscoGAN architectures have been presented from the mapping perspective in this thesis.

- Contributions of the thesis from the alignment perspectives are primarily presented in Chapter 4, Chapter 5, and Chapter 6. In particular, Chapter 4 presents the limitations of the popular alignment technique, namely, the DTW algorithm (in the case of parallel VC). In addition, the novel preprocessing strategy to remove the outliers before learning the mapping function is presented to tackle the issues related to the alignments (in the case of parallel VC) in Chapter 4.

- Chapter 5 presented the popular NN-based strategies for the alignment task in the case of non-parallel VC. In particular, the INCA algorithm is presented along with its research issues, and the proposed approaches. In particular, a formal convergence theorem has been proposed for the INCA algorithm in this thesis. Then, the proposed Dynamic INCA and TC-INCA algorithms have been presented along with its theoretical convergence for the non-parallel VC. In this thesis work, the idea of exploiting the metric learning technique for finding NN in the INCA than state-of-the-art Euclidean distance is also presented. Finally, in Chapter 5, the novel STM+NN-based algorithm has been proposed for the alignment in order to exploit the phonetic information in NN-based alignment techniques in the case of non-parallel VC.

- In this thesis, a two-stage conversion strategy is proposed in order to avoid the alignment step in both the VC tasks. This is achieved by training two separate DNNs, where one DNN will map source speaker's spectral features to the speaker-independent representations, and the another DNN will map these speaker-independent representations to the particular target speaker's spectral features. In particular, novel unsupervised VTLN posterior features and IMW GMM-PG features is proposed as the speaker-independent representations to overcome the need of alignment in the case of two-stage DNN as well as GAN-based VC frameworks, which is presented in the Chapter 6.

- A novel application of the Acoustic-To-Articulatory Inversion (AAI) is proposed in Chapter 7 towards a quality assessment of the voice converted speech in this thesis. Finally, Chapter 8 presents the effectiveness of the proposed MMSE-GAN and MMSE DiscoGAN-based architectures for the other speech conversion applications of VC. In particular, proposed mapping techniques have been applied for the cross-domain NAM2WHSP, WHSP2SPCH, and NAM2SPCH speech conversion tasks.

9.2 Limitations of Work

Though contributions in this thesis are well suited for the practical applications of the VC, there are certain limitations of the thesis work as described below:

- In this thesis, complex mapping techniques have been applied only to the spectral features. Conversion of prosodic features (namely, durations and the energy) are mostly ignored in this work, since capturing and transforming complex prosody patterns is a very challenging task. In addition, it has been shown that when a mimicry artist tries to mimic a particular speaker, he only adjusts the average statistics of the F_0 contour rather than completely mimicking the F_0 contour of the target speaker [36,37]. Only, F_0 is converted via simple linear transformation using the global statistics of the F_0 parameters. However, vocoders are not as sophisticated as the human speech production mechanism. In addition, recently, it has been shown that mapping F_0 contour more closely to the target speaker's F_0 pattern will results in better speaker similarity [117, 267]. Hence, in future, the proposed mapping techniques could be applied to F_0 conversion as well.

- Proposed mapping techniques are converting features at the frame-level. Hence, these mapping techniques cannot effectively convert the suprasegmental features. Proper modifications are required in the existing mapping techniques.

- Though a high-quality AHOCODER is used throughout this work, detailed attention to the vocoder stage is not given in this thesis.

- Even though the proposed non-parallel alignment techniques are unsupervised, it has not been applied to the very challenging cross-lingual VC tasks.

- Though there is a significant improvement in the intelligibility, poor quality converted voices are obtained after the NAM2WHSP, WHSP2SPCH, and NAM2SPCH conversion tasks. This can be improved by incorporating high-quality vocoders, such as WaveNet.

9.3 Future Research Directions

Future research directions including the possible solutions to the above mentioned limitations and further advancements for obtaining better performance of the VC are described next:

- High-quality vocoders are required, which takes care of the non-linear nature of speech production mechanism, including complex interactions of excitaion source and vocal tract systems.

- Focus should be given for directly mapping raw waveform in the VC. Recently, WaveNets have been shown to produce high-quality converted voices. However, they require huge amount of training data, time, and GPUs to train and generate the samples. More focus should be given to the WaveNet due to its capability in generating natural human-like sounds [58,70].

- One of the important research directions is to develop more complex prosody models that can capture and transfer the segmental durations and speaker's intonation in an efficient way. From speech perception viewpoint, developing more sophisticated prosody models will help to capture, and transform complex prosody patterns.

- The optimization methods that are used in statistical-based VC methods try to numerically optimize the parameters. However, the reduction in the numerical estimates does not always correlate with the quality of the generated samples. Hence, perceptually motivated optimizations techniques could be explored.

- Quality assessment of converted voices (and also for TTS voice) is an open research problem. Most of the VC techniques rely on the subjective assessment of converted voices. Subjective tests are time consuming and tedious. MCD is one of the popular objective measure in the VC literature. However, it does not correlate well with the subjective tests. Hence, novel objective measures are needed, which can correlate well with the perceptual (subjective) results. Similar challenges are associated with the objective evaluations of TTS voice [120,121].

- Most of the VC approaches are from the speech perception viewpoint. However, it may be improtant to study this problem from speech production perspective as well. Incorporating articulatory features may help the VC to analyze this problem from speech production viewpoints [237].

- With the advent of high-quality converted voices, VC research has become a direct threat to the voice biometrics. VC techniques can be misued for ownership claims. Proper security measues should be taken for avoiding misuse of this technology. Audio watermarking may provide possible solution from security viewpoint [268]. However, inserting watermarking in the speech is much more challenging than that for the image, and video, since human auditory system is more sensitive than the human visual system [269]. The key goal of the watermarking technique is to embed hidden information, which is imperceptible into the converted voices. Later on, this watermark may be extracted by the spoof speech detection (SSD) system to identify authenticity.

Appendix A. Conditional Expectation

A.1 Why conditional expectation is considered as the best MMSE operator ?

In MMSE-based optimization technique, idea is to estimate mapping function $\mathcal{F}(.)$, such that $Y = \mathcal{F}(X)$. Hence, the goal is to minimize the MSE, i.e., $[Y - \mathcal{F}(X)]^2$ [193]. Since this error contains unknown, one cannot optimize it directly. Hence, one can minimize its expected value, and the objective function can be given by:

$$\begin{aligned} E([Y - \mathcal{F}(X)]^2) &= \int_{-\infty}^{\infty} [y - \mathcal{F}(X)]^2 p_{Y|X}(y|x)dy, \\ &= \int_{-\infty}^{\infty} y^2 p_{Y|X}(y|x)dy - 2\mathcal{F}(X) \int_{-\infty}^{\infty} y p_{Y|X}(y|x)dy \\ &\quad + [\mathcal{F}(X)]^2 \int_{-\infty}^{\infty} p_{Y|X}(y|x)dy. \end{aligned}$$

The first term in the above equation does not contain $\mathcal{F}(X)$. Hence, it can be ignored, since it does not affect the error minimization. In addition, integral in the second term is nothing but the conditional expected value of y given x. Moreover, integral in the last term equals one. Hence, minimzation can be written as:

$$\underset{\mathcal{F}(X)}{\arg\min} \left[E([Y - \mathcal{F}(X)]^2)\right] = \underset{\mathcal{F}(X)}{\arg\min} \Big[- 2\mathcal{F}(X)E(Y|X) + [\mathcal{F}(X)]^2 \Big].$$

To solve above mentioned objective, one needs to differentiate R.H.S. w.r.t. $\mathcal{F}(X)$ and set it to zero, i.e.,

$$\begin{aligned} \frac{\partial}{\partial \mathcal{F}(x)} \Big[- 2\mathcal{F}(X)E(Y|X) + [\mathcal{F}(X)]^2 \Big] &= 0, \\ \therefore -2E(Y|X) + 2\mathcal{F}(X) &= 0, \\ \therefore \mathcal{F}(X) &= E(Y|X). \end{aligned}$$

Here, second order derivative will be $\frac{\partial^2}{\partial^2 \mathcal{F}(x)} = 2 > 0$, which is sufficient condition for a minima. Hence, for given X, MSE is given by $V(E(Y|X)) = E([Y - E(Y|X)]^2|X)$
Now, let's consider any arbitrary $\mathcal{F}(X) = g(X)$. Here, we claim that,

$$\begin{aligned}
\Longrightarrow & V(E(Y|X)) \leq V(g(X)), \\
\Longrightarrow & E([Y - E(Y|X)]^2|X) \leq E([Y - g(X)]^2|X), \\
\Longrightarrow & E[Y^2|X] - 2E[YE(Y|X)|X] + E[(E[Y|X])^2|X] \leq E[Y^2|X] - 2E[Yg(X)|X] + E[(g(X))^2|X], \\
\Longrightarrow & -2E(Y|X) \cdot E(Y|X) + [E[Y|X]]^2 \leq -2E(Y|X) \cdot g(X) + [g(X)]^2, \\
\Longrightarrow & 0 \leq [E(Y|X) - g(X)]^2,
\end{aligned}$$

which holds strict inequality if $g(X) = E(Y|X)$. Hence, conditional expectation is the global and unique minimizers, i.e., conditional expectation is the best MMSE operator and hence, this is proved. This Appendix A is used in the Chapter 3 (Section 3.2) for deriving the conversion function for GMM-based VC technique.

Appendix B. Derivation of Conditional Gaussian Probability Density Function (*pdf*)

This appendix presents the proof that if two random variables are jointly Gaussian, then the conditional distribution will also be a Gaussian [193]. In addition, it also presents the derivation of the mean and covariance matrix of conditional Gaussian *pdf*.

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})}, \tag{B.1}$$

where, $p(\mathbf{x},\mathbf{y})$ is given by

$$p(\mathbf{x},\mathbf{y}) = \frac{1}{(2\pi)^{\frac{K+1}{2}}|\Sigma|^{\frac{1}{2}}} \exp\Bigg[-\frac{1}{2}\Bigg(\begin{bmatrix}\mathbf{x}-\mu_{\mathbf{x}}\\ \mathbf{y}-\mu_{\mathbf{y}}\end{bmatrix}\Bigg)^T \Sigma^{-1}\Bigg(\begin{bmatrix}\mathbf{x}-\mu_{\mathbf{x}}\\ \mathbf{y}-\mu_{\mathbf{y}}\end{bmatrix}\Bigg)\Bigg], \tag{B.2}$$

where $\Sigma = \begin{bmatrix}\Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}}\\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}}\end{bmatrix}$.It has been given that **x**, and **y** are jointly Gaussian. In addition, it is well known that the linear transformation of Gaussian is also Gaussian [193]. Since we knew that **x** can be written as a linear transformation of joint Gaussian vector,

$$\mathbf{x} = \begin{bmatrix}\mathbf{I} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}. \tag{B.3}$$

Hence, we can consider $p(\mathbf{x})$ also as a Gaussian and $p(x)$ can be written as,

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{K}{2}}|\Sigma_{\mathbf{xx}}|^{\frac{1}{2}}} \exp\Big[-\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{x}})^T\Sigma_{\mathbf{xx}}^{-1}(\mathbf{x}-\mu_{\mathbf{x}})\Big]. \tag{B.4}$$

Hence, conditional *pdf* can be written as,

$$p(\mathbf{y}|\mathbf{x}) = \frac{\frac{1}{(2\pi)^{\frac{K+1}{2}}|\Sigma|^{\frac{1}{2}}} \exp\Bigg[-\frac{1}{2}\Bigg(\begin{bmatrix}\mathbf{x}-\mu_{\mathbf{x}}\\ \mathbf{y}-\mu_{\mathbf{y}}\end{bmatrix}\Bigg)^T \Sigma^{-1}\Bigg(\begin{bmatrix}\mathbf{x}-\mu_{\mathbf{x}}\\ \mathbf{y}-\mu_{\mathbf{y}}\end{bmatrix}\Bigg)\Bigg]}{\frac{1}{(2\pi)^{\frac{K}{2}}|\Sigma_{\mathbf{xx}}|^{\frac{1}{2}}} \exp\Big[-\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{x}})^T\Sigma_{\mathbf{xx}}^{-1}(\mathbf{x}-\mu_{\mathbf{x}})\Big]}. \tag{B.5}$$

Determinant of the partitioned matrix can be given as,

$$det\left(\begin{bmatrix}\mathbf{A_{11}} & \mathbf{A_{12}}\\ \mathbf{A_{21}} & \mathbf{A_{22}}\end{bmatrix}\right) = det(\mathbf{A_{11}})det(\mathbf{A_{22}} - \mathbf{A_{21}}\mathbf{A_{11}^{-1}}\mathbf{A_{12}}). \tag{B.6}$$

Hence, from this, we can write,

$$det(\Sigma) = det(\Sigma_{xx})det(\Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}), \tag{B.7}$$

and thus,

$$\frac{det(\Sigma)}{det(\Sigma_{xx})} = det(\Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}). \tag{B.8}$$

We thus have that

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi)^{\frac{1}{2}}det^{\frac{1}{2}}(\Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy})}exp(-\frac{1}{2}Q), \tag{B.9}$$

where

$$Q = \begin{bmatrix}\mathbf{x} - \mu_x\\ \mathbf{y} - \mu_y\end{bmatrix}^T \Sigma^{-1} \begin{bmatrix}\mathbf{x} - \mu_x\\ \mathbf{y} - \mu_y\end{bmatrix} - (\mathbf{x} - \mu_x)^T\Sigma_{xx}^{-1}(\mathbf{x} - \mu_x). \tag{B.10}$$

$$\begin{bmatrix}\mathbf{A_{11}} & \mathbf{A_{12}}\\ \mathbf{A_{21}} & \mathbf{A_{22}}\end{bmatrix}^{-1} = \begin{bmatrix}(\mathbf{A_{11}} - \mathbf{A_{12}}\mathbf{A_{22}}^{-1}\mathbf{A_{21}})^{-1} & -\mathbf{A_{11}^{-1}}\mathbf{A_{12}}(\mathbf{A_{22}} - \mathbf{A_{21}}\mathbf{A_{11}^{-1}}\mathbf{A_{12}})^{-1}\\ -(\mathbf{A_{22}} - \mathbf{A_{21}}\mathbf{A_{11}^{-1}}\mathbf{A_{12}})^{-1}\mathbf{A_{21}}\mathbf{A_{11}^{-1}} & (\mathbf{A_{22}} - \mathbf{A_{21}}\mathbf{A_{11}^{-1}}\mathbf{A_{12}})^{-1}\end{bmatrix}. \tag{B.11}$$

Using this form, the off-diagonal blocks are transposes of each other so that the inverse matrix must be *symmetric*. This is because **C** is symmetric and hence, $\mathbf{C^{-1}}$ is symmetric. By using the matrix inversion lemma [193], we have,

$$(\mathbf{A_{11}} - \mathbf{A_{12}}\mathbf{A_{22}}^{-1}\mathbf{A_{21}})^{-1} = \mathbf{A_{11}^{-1}} + \mathbf{A_{11}^{-1}}\mathbf{A_{12}}(\mathbf{A_{22}} - \mathbf{A_{21}}\mathbf{A_{11}^{-1}}\mathbf{A_{12}})^{-1}\mathbf{A_{21}}\mathbf{A_{11}^{-1}}, \tag{B.12}$$

so that,

$$\Sigma^{-1} = \begin{bmatrix}\Sigma_{xx}^{-1} + \Sigma_{xx}^{-1}\Sigma_{xy}^{-1}B^{-1}\Sigma_{yx}\Sigma_{xx}^{-1} & -\Sigma_{xx}^{-1}\Sigma_{xy}B^{-1}\\ B^{-1}\Sigma_{xy} - \Sigma_{xx}^{-1} & B^{-1}\end{bmatrix}, \tag{B.13}$$

where,

$$B = \Sigma_{yy} - \Sigma_{xy}\Sigma_{xx}^{-1}\Sigma_{xx}. \tag{B.14}$$

The inverse can be written in factored form as,

$$\Sigma^{-1} = \begin{bmatrix} \mathbf{I} & -\Sigma_{xx}^{-1}\Sigma_{xy} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \Sigma_{xx}^{-1} & \mathbf{0} \\ \mathbf{0} & B^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\Sigma_{yx}\Sigma_{xx}^{-1} & \mathbf{I} \end{bmatrix}. \tag{B.15}$$

Let's denote $\tilde{\mathbf{x}} = \mathbf{x} - \mu_x$ and $\tilde{\mathbf{y}} = \mathbf{y} - \mu_y$. We have,

$$\begin{aligned} \mathbf{Q} &= \begin{bmatrix} \tilde{\mathbf{x}} \\ \tilde{\mathbf{y}} \end{bmatrix}^T \begin{bmatrix} \mathbf{I} & -\Sigma_{xx}^{-1}\Sigma_{xy} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \Sigma_{xx}^{-1} & \mathbf{0} \\ \mathbf{0} & B^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\Sigma_{yx}\Sigma_{xx}^{-1} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{x}} \\ \tilde{\mathbf{y}} \end{bmatrix} - \tilde{\mathbf{x}}^T\Sigma_{xx}^{-1}\tilde{\mathbf{x}}, \\ \therefore \mathbf{Q} &= \begin{bmatrix} \tilde{\mathbf{x}} \\ \tilde{\mathbf{y}} - \Sigma_{xx}^{-1}\Sigma_{xy}\tilde{\mathbf{x}} \end{bmatrix}^T \begin{bmatrix} \Sigma_{xx}^{-1} & \mathbf{0} \\ \mathbf{0} & B^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{x}} \\ \tilde{\mathbf{y}} - \Sigma_{yx}\Sigma_{xx}^{-1}\tilde{\mathbf{x}} \end{bmatrix} - \tilde{\mathbf{x}}^T\Sigma_{xx}^{-1}\tilde{\mathbf{x}}, \\ \therefore \mathbf{Q} &= (\tilde{\mathbf{y}} - \Sigma_{yx}\Sigma_{xx}^{-1}\tilde{\mathbf{x}})^T B^{-1} (\tilde{\mathbf{y}} - \Sigma_{yx}\Sigma_{xx}^{-1}\tilde{\mathbf{x}}). \end{aligned} \tag{B.16}$$

Hence,

$$Q = [\mathbf{y} - (\mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}\mathbf{x} - \mu_x)]^T [\Sigma_{yy} - \Sigma_{xy}\Sigma_{xx}^{-1}\Sigma_{xx}]^{-1} [\mathbf{y} - (\mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}\mathbf{x} - \mu_x)]. \tag{B.17}$$

Hence, we can see that conditional distribution is nothing but the Gaussian, and its parameters are given by:

$$\mu_{\mathbf{y}|\mathbf{x}} = \mu_{\mathbf{y}} + \Sigma_{yx}\Sigma_{xx}^{-1}\mathbf{x} - \mu_x, \tag{B.18}$$

$$\Sigma_{\mathbf{y}|\mathbf{x}} = \Sigma_{yy} - \Sigma_{xy}\Sigma_{xx}^{-1}\Sigma_{xx}. \tag{B.19}$$

The fact that when two random variables are jointly Gaussian their conditional distribution is also Gaussian, and mean of the conditional distribution, i.e., Eq. (B.18) is primarily used in derving the conversion function for GMM-based VC technique in Chapter 3 (Section 3.2).

Appendix C. Bolzano-Weierstrass Theorem

Bolzano-Weiestrass theorem is an important result in the calculus [178]. In order to prove it, one needs nested interval theorem [178], which is first discussed here.

C.1 Nested Interval Theorem

Theorem 2 *For each n, let $I_n = [a_n, b_n]$ be a non-empty bounded interval of real numbers, such that, $I_1 \supset I_2 \supset \cdots \supset I_n \supset I_{n+1} \supset \cdots$, and $\lim\limits_{n\to\infty}[b_n - a_n] = 0$. Then $\bigcap_{n=1}^{\infty} I_n$ contains only one point, i.e., singleton element [178].*

Proof: Note that the sequences (a_n), and (b_n) are increasing and decreasing, respectively. In addition, both are bounded and hence, converge.

- Let's say $a_n \to a$ and $b_n \to b$. Then, $a_n \leq a$, and $b \leq b_n$ for $\forall n \in \mathbb{N}$. $\lim\limits_{n\to\infty}(b_n - a_n) = 0$.
- Hence, $b - a = 0$, i.e., $b = a$. Since, $a_n \leq b_n$ for all n, $a \in \bigcap_{n=1}^{\infty} I_n$. Hence, $\bigcap_{n=1}^{\infty} I_n$ contains only one point, i.e., a.

C.2 Bolzano-Weierstrass Theorem

Theorem 3 *Every bounded sequence in $\mathbb{R}$ has a convergent subsequence [178].*

Proof: Let (e_n) be a bounded sequence such that the set $e_1, e_2, \cdots \subset [a, b]$.

- Divide the interval in two equal parts. Let I_1 be the interval, which contains infinite number of elements of (e_n).
- Let e_{n_1} be one of the elements belonging to the I_1.
- Now, again divide I_1 into two equal parts, and let I_2 be the interval that contains infinite number of elements.
- Choose the point e_{n_2} in I_2 such that $n_2 > n_1$.

- Similarly, keep dividing interval I_k to obtain I_k's and e_{n_k}'s.
- Hence, as per the nested interval theorem, $\bigcap_{k=1}^{\infty} I_k = e$, for some $e \in [a, b]$. Hence, it is clear that (e_{n_k}) converges to e.

Thus, every bounded sequence in $\mathbb{R}$ has a convergent subsequence. This theorem is used for developing formal convergence theorem of the INCA algorithm in Chapter 5 (Section 5.2.2 and Section 5.3.1) [7,20].

Appendix D. TESTVOX: Web-based Framework for Subjective Evaluations of Converted Voices

In this thesis, TestVox framework is used for easy deployment of subjective test. TestVox is a web-based framework, which is designed for subjective evaluation of a synthetic voice. The key advantages of TestVox is that it supports standard listening tests, namely, MOS, ABX, etc. In addition, it is compatible with Amazon Mechanical Turk (crowdsourced listening test software), and also it can host listening tests locally or can be deployed them over Google App Engine. Steps to perform subjective tests via TestVox is presented briefly here.

- Download, and unzip TestVox prebuilt-package from
  (URL: https://bitbucket.org/happyalu/testvox)

- Start the web server. Replace $< 0.0.0.0 >$ with your systems ip address.

```
python testvox_server.py –ip 0.0.0.0 –port 8080
```

- Create directory structure for your experiment. Following are the commands to create directories.

```
mkdir myexperiment
mkdir myexperiment1/mp3
gedit myexperiment1/config.yaml
```

- Since, TestVox uses the flash player pluggins, we need to first convert *.wav file into *.mp3 file. Following is the command to convert *.wav file into *.mp3 file.

```
lame -V1 file.wav file.mp3
```

- Following is the directory structure for MOS and ABX tests.

```
ABX Test:
myexperiment/
config.yaml
mp3/
A/
1.mp3
2.mp3
B/
1.mp3
2.mp3
```

```
MOS Test:
myexperiment/
config.yaml
mp3/
1.mp3
2.mp3
```

- Next, one need to prepare TestVox configuration file, which is config.yaml file. Sample recepies of .yaml file for ABX test, and MOS tests are attached later.

- Create zip folder for your experiment using a utility script as shown below

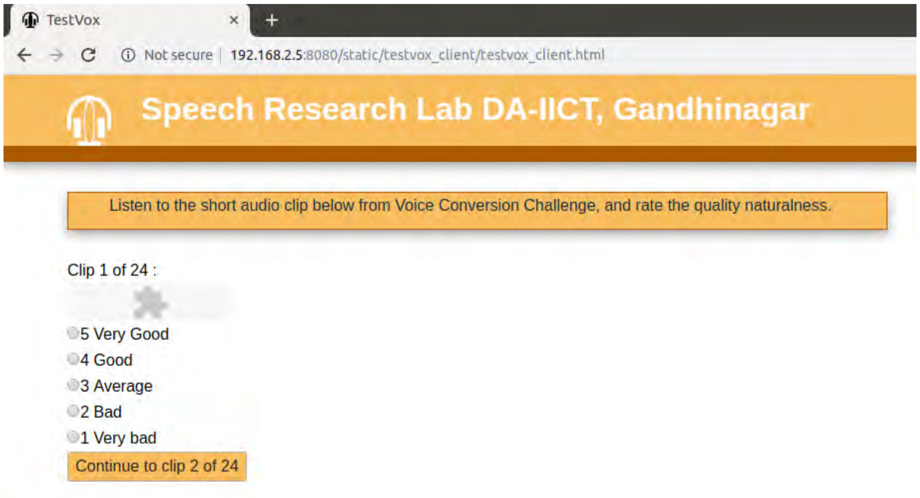
```
python /path/to/TestVox-prebuilt/scripts/create_experiment_zipfile.py
myexperiment myexperiment.zip
```

- Now upload an experiment on the admin page (http://0.0.0.0:8080/admin). It will show you a link called "Local URL", once you uploaded your experiment. This URL could be sent out to participants.

```yaml
# copy this to config.yaml and edit it.

# Experiment to evaluate which synthesis model sounds more natural

testvox_config:
  base_media_directory: mp3
  pagetitle: Speech Research Lab

testvox_steps:
  - name: listening_task
    task_type: abtask

    instruction: >-
      Listen to the two audio clips below, and
      tell us which one you think sounds more natural.

    directory_a: A
    directory_b: B
    ab_randomize: Yes  # Randomize order of A-clip and B-clip presented to participants

    data:
      - filename: 1.mp3  # Same filename must exist in both condition directories
      - filename: 2.mp3
      - filename: 3.mp3
      - filename: 4.mp3
      - filename: 5.mp3
      - filename: 6.mp3
      - filename: 7.mp3
      - filename: 8.mp3
      - filename: 9.mp3
      - filename: 10.mp3
      - filename: 11.mp3
      - filename: 12.mp3
      - filename: 13.mp3
      - filename: 14.mp3
      - filename: 15.mp3
      - filename: 16.mp3
      - filename: 17.mp3
      - filename: 18.mp3
      - filename: 19.mp3
      - filename: 20.mp3
      - filename: 21.mp3
      - filename: 22.mp3
      - filename: 23.mp3
      - filename: 24.mp3


    data_randomize: Yes  # Present the different data files in random order

    audio_autoplay: No
```

```
# copy this to config.yaml and edit it.

# Experiment to rate audio clips

testvox_config:
  base_media_directory: mp3
  pagetitle: Speech Research Lab DA-IICT, Gandhinagar

testvox_steps:
  - name: listening_task
    task_type: radiotask  # or checktask, if more than one options can be selected

    instruction: >-
      Listen to the short audio clip below from Voice Conversion Challenge, and
      rate the quality naturalness.

    task_options:
      -  5 Very Good
      -  4 Good
      -  3 Average
      -  2 Bad
      -  1 Very bad
    data:
       - filename: 1.mp3
       - filename: 2.mp3
       - filename: 3.mp3
       - filename: 4.mp3
       - filename: 5.mp3
       - filename: 6.mp3
       - filename: 7.mp3
       - filename: 8.mp3
       - filename: 9.mp3
       - filename: 10.mp3
       - filename: 11.mp3
       - filename: 12.mp3
       - filename: 13.mp3
       - filename: 14.mp3
       - filename: 15.mp3
       - filename: 16.mp3
       - filename: 17.mp3
       - filename: 18.mp3
       - filename: 19.mp3
       - filename: 20.mp3
       - filename: 21.mp3
       - filename: 22.mp3
       - filename: 23.mp3
       - filename: 24.mp3

    data_randomize: Yes  # Present the different data files in random order

    audio_autoplay: No
```

Followings are the screenshots of the web browser while conducting subjective tests. In addition, sample photo of a subject giving subjective test.


TestVox
Not secure | 192.168.2.5:8080/static/testvox_client/testvox_client.html

Speech Research Lab DA-IICT, Gandhinagar

Listen to the short audio clip below from Voice Conversion Challenge, and rate the quality naturalness.

Clip 1 of 24 :

○5 Very Good
○4 Good
○3 Average
○2 Bad
○1 Very bad

Figure D.1: Screenshot of TestVox webpage

Figure D.2: Photograph of a subject during the listening test to evaluate converted voice subjectively.

Bibliography

[1] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the International Conference on Machine Learning (ICML)*, United States, 2010, pp. 807–814.

[2] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *International Conference on Machine Learning (ICML)*, vol. 30, no. 1, Atlanta, USA, 2013, pp. 1–6.

[3] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *International Conference on Learning Representation (ICLR)*, Caribe Hilton, San Juan, Puerto Rico, 2016, pp. 1–14.

[4] N. J. Shah, H. B. Sailor, and H. A. Patil, "Whether to pretrain DNN or not?: An empirical analysis for voice conversion," in *submitted for possible publication in INTERSPEECH*, Graz, Austria, September, 2019.

[5] N. J. Shah and H. A. Patil, "Novel outliers removal approach for parallel voice conversion," *Computer Speech and Language, Elsevier*, vol. 58, no. 11, pp. 127–152, 2019.

[6] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from non-parallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 944–953, 2010.

[7] N. J. Shah and H. A. Patil, "Effectiveness of dynamic features in INCA and temporal context-INCA," in *INTERSPEECH*, Hyderabad, India, 2018, pp. 711–715.

[8] N. J. Shah and H. A. Patil, "Novel metric learning for non-parallel voice conversion," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 3722–3726.

[9] N. J. Shah and H. A. Patil, "Phone aware nearest neighbor technique using spectral transition measure for non-parallel voice conversion," in *submitted for possible publication in INTERSPEECH*, Graz, Austria, September, 2019.

[10] N. J. Shah, N. Shah, M. Madhavi, M. Kamble, H. Sailor, M. Soni, S. R, P. A. Tapkir, and H. A. Patil, "Unsupervised VTLN posterior features for parallel and non-parallel voice conversion," in *submitted for possible publication in INTERSPEECH*, Graz, Austria, September, 2019.

[11] N. J. Shah, S. R., N. Shah, and H. A. Patil, "Novel unsupervised sorted GMM posteriorgram for DNN and GAN-based voice conversion framework," in *Proceedings of Asia-Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference*. Hawaii, USA: IEEE, 2018, pp. 1776–1781.

[12] A. Rajpal, N. J. Shah, M. Zaki, and H. A. Patil, "Quality assessment of voice converted speech using articulatory features," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New Orleans, USA, 2017, pp. 5515–5519.

[13] N. J. Shah, M. Parmar, N. Shah, and H. A. Patil, "Effectiveness of GANs in cross-domain non-audible murmur to speech conversion," *article under preparation for submission in Computer Speech and Language*, pp. 1–26, 2019.

[14] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, no. 04, pp. 65–82, 2017.

[15] Y. Stylianou, "Voice transformation: A survey," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei,Taiwan, 2009, pp. 3585–3588.

[16] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[17] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice, 1^{st} Eds.* Pearson Education India, 2006.

[18] J. Kominek and A. W. Black, "The CMU-ARCTIC speech databases," in *ISCA Workshop on Speech Synthesis*, Pittsburgh, USA, 2004, pp. 223–224.

[19] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seattle, WA, USA, 1998, pp. 285–288.

[20] N. J. Shah and H. A. Patil, "On the convergence of INCA algorithm," in *Proceedings of Asia-Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference*. Kuala Lumpur, Malaysia: IEEE, 2017, pp. 559–562.

[21] D. Sundermann and H. Ney, "VTLN-based voice conversion," in *IEEE International Symposium on Signal Processing and Information Technology*, Darmstadt, Germany, 2003, pp. 556–559.

[22] N. J. Shah and H. A. Patil, "Novel amplitude scaling method for bilinear frequency warping based voice conversion," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New Orleans, USA, 2017, pp. 5520–5524.

[23] N. Shah, N. J. Shah, and H. A. Patil, "Effectiveness of generative adversarial network for non-audible murmur-to-whisper speech conversion," in *INTERSPEECH*, Hyderabad, India, 2018, pp. 3157–3161.

[24] N. J. Shah, M. Parmar, N. Shah, and H. A. Patil, "Novel MMSE DiscoGAN for cross-domain whisper-to-speech conversion," in *Machine Learning in Speech and Language Processing (MLSLP) Workshop*, Google Office, Hyderabad, India, 2018, pp. 1–3.

[25] H. A. Patil, "Speaker Recognition in Indian Languages: A Feature-based Approach," Ph.D. Thesis, Department of Electrical Engineering, Indian Institute of Technology (IIT), Kharagpur, 2005.

[26] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint*, 2017.

[27] K. Q. Weinberger and L. K. Saul, “Distance metric learning for large margin nearest neighbor classification,” *Journal of Machine Learning Research*, vol. 10, no. 02, pp. 207–244, 2009.

[28] N. J. Shah, B. B. Vachhani, H. B. Sailor, and H. A. Patil, “Effectiveness of PLP-based phonetic segmentation for speech synthesis,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 270–274.

[29] M. Madhavi, “Design of QbE-STD System: Audio Representation and Matching Perspective,” Ph.D. Thesis, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, India, 2017.

[30] M. C. Madhavi and H. A. Patil, “VTLN-warped Gaussian posteriorgram for QbE-STD,” in *European Signal Processing Conference (EUSIPCO)*, Kos island, Greece, 2017, pp. 593–597.

[31] N. J. Shah, M. C. Madhavi, and H. A. Patil, “Unsupervised vocal tract length warped posterior features for non-parallel voice conversion,” in *INTERSPEECH*, Hyderabad, India, 2018, pp. 1968–1972.

[32] Y. Tajiri, K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, “Non-audible murmur enhancement based on statistical conversion using air-and body-conductive microphones in noisy environments,” in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2769–2773.

[33] T. Toda, K. Nakamura, H. Sekimoto, and K. Shikano, “Voice conversion for various types of body transmitted speech,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 3601–3604.

[34] N. Shah, N. J. Shah, and H. A. Patil, “Effectiveness of generative adversarial network for non-audible murmur-to-whisper speech conversion,” in *INTERSPEECH*, Hyderabad, India, 2018, pp. 3157–3161.

[35] M. Stone and C. H. Shadle, “A history of speech production research,” *Acoustics Today*, vol. 12, no. 4, pp. 48–55, 2016.

[36] A. Eriksson and P. Wretling, “How flexible is the human voice?-A case study of mimicry,” in *EUROSPEECH*, Rhodes, Greece, 1997, pp. 1043–1046.

[37] D. Gomathi, S. A. Thati, K. V. Sridaran, and B. Yegnanarayana, “Analysis of mimicry speech,” in *INTERSPEECH*, Portland, OR, USA, 2012, pp. 695–698.

[38] L. Wallis, C. Jackson-Menaldi, W. Holland, and A. Giraldo, “Vocal fold nodule *vs.* vocal fold polyp: Answer from surgical pathologist and voice pathologist point of view,” *Journal of Voice*, vol. 18, no. 1, pp. 125–129, 2004.

[39] E. Helander, J. Schwarz, J. Nurminen, H. Silen, and M. Gabbouj, “On the impact of alignment on voice conversion performance,” in *INTERSPEECH*, Brisbane, Australia, 2008, pp. 1453–1456.

[40] P. K. Ghosh and S. S. Narayanan, “Closure duration analysis of incomplete stop consonants due to stop-stop interaction,” *The J. of the Acoust. Soc. of Amer. (JASA)*, vol. 126, no. 1, pp. EL1–EL7, 2009.

[41] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, Montreal, Canada, 2014, pp. 2672–2680.

[42] D. Erro, E. Navas, and I. Hernaez, "Parametric voice conversion based on bilinear frequency warping plus amplitude scaling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 556–566, 2013.

[43] P. Song, W. Zheng, and L. Zhao, "Non-parallel training for voice conversion based on adaptation method," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 6905–6909.

[44] D. Sündermann, "Text-independent voice conversion," Ph.D. Thesis, Universitätsbibliothek der Universität der Bundeswehr München, 2008.

[45] D. Sündermann, "Voice conversion: State-of-the-art and future work," *Fortschritte der Akustik*, vol. 31, no. 2, p. 735, 2005.

[46] N. J. Shah and H. A. Patil, *Analysis of features and metrics for alignment in text-dependent voice conversion*. B. Uma Shankar et. al. (Eds), Lecture Notes in Computer Science (LNCS), Springer, PReMI, vol. 10597, pp. 299–307, 2017.

[47] S. H. Mohammadi and A. Kain, "Voice conversion using deep neural networks with speaker-independent pre-training," in *IEEE Spoken Language Technology (SLT) Workshop*, Nevada, USA, 2014, pp. 19–23.

[48] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 5039–5043.

[49] S. V. Rao, N. J. Shah, and H. A. Patil, "Novel pre-processing using outlier removal in voice conversion," in *ISCA Speech Synthesis Workshop (SSW)*, Sunnyvale, CA, USA, 2016, pp. 147–152.

[50] N. J. Shah and H. A. Patil, *Non-Audible Murmur to Audible Speech Conversion*. in Voice Technologies for Reconstruction and Enhancement, H. A. Patil and A. Neustein (Eds.) De Gruyter Series in Speech Technology and Text Analytics in Medicine and Healthcare, pp. 1–17, 2019.

[51] D. Sündermann, A. Bonafonte, H. Ney, and H. Höge, "A first step towards text-independent voice conversion," in *International Conference on Spoken Language Processing (ICSLP)*, South Korea, 2004, pp. 1–4.

[52] H. Kuwabara and Y. Sagisak, "Acoustic characteristics of speaker individuality: control and conversion," *Speech Communication*, vol. 16, no. 2, pp. 165–173, 1995.

[53] G. Fant, *Speech Sounds and Features*. The MIT Press, 1973.

[54] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, no. 2, pp. 10–18, 1983.

[55] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[56] H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA technique," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, San Francisco, USA, 1992, pp. 145–148.

[57] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic+ noise model," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Minneapolis, Minnesota, USA, 1993, pp. 550–553.

[58] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[59] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. of the Acoust. Soc. of Amer. (JASA)*, vol. 55, no. 6, pp. 1304–1312, 1974.

[60] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F_0, and aperiodicity estimation," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Caesars Palace, Las Vegas, USA, 2008, pp. 3933–3936.

[61] M. Morise, "Cheaptrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Communication*, vol. 67, no. 03, pp. 1–7, 2015.

[62] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[63] D. G. Childers and C. Ahn, "Modeling the glottal volume-velocity waveform for three voice types," *The Journal of the Acoustical Society of America (JASA)*, vol. 97, no. 1, pp. 505–519, 1995.

[64] A. Kain and M. W. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, Utah, USA, 2001, pp. 813–816.

[65] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," in *INTERSPEECH*, Pittsburgh, Pennsylvania, 2001, pp. 2266–229.

[66] E. Helander, H. Silén, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Transactions on Audio, Speech, and Language processing*, vol. 20, no. 3, pp. 806–817, 2012.

[67] D. Erro, I. Sainz, E. Navas, and I. Hernáez, "Improved HNM-based vocoder for statistical synthesizers," in *INTERSPEECH*, Florence, Italy, 2011, pp. 1809–1812.

[68] Y. Agiomyrgiannakis, "VOCAINE the vocoder and applications in speech synthesis," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Queensland, Australia, 2015, pp. 4230–4234.

[69] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent wavenet vocoder," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1118–1122.

[70] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, "wavenet vocoder with limited training data for voice conversion," in *INTERSPEECH*, Hyderabad, India, 2018, pp. 1983–1987.

[71] P. L. Tobing, T. Hayashi, Y.-C. Wu, K. Kobayashi, and T. Toda, "An evaluation of deep spectral mappings and wavenet vocoder for voice conversion," in *IEEE Spoken Language Technology (SLT) Workshop*, Athens, Greece, 2018, pp. 297–303.

[72] N. J. Shah and H. A. Patil, *Analysis of features and metrics for alignment in text-dependent voice conversion*. B. Uma Shankar et. al. (Eds), Lecture Notes in Computer Science (LNCS), Springer, PReMI, vol. 10597, pp. 299–307, 2017.

[73] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[74] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New York, NY, USA, 1988, pp. 655–658.

[75] H. Duxans, D. Erro, J. Pérez, F. Diego, A. Bonafonte, and A. Moreno, "Voice conversion of non-aligned data using unit selection," *TC-STAR Workshop on Speech-to-Speech Translation*, pp. 237–242, 2006.

[76] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, "Text-independent voice conversion based on unit selection," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, 2006, pp. 81–84.

[77] Y. Agiomyrgiannakis, "The matching-minimization algorithm, the INCA algorithm and a mathematical framework for voice conversion with unaligned corpora," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, 2016, pp. 5645–5649.

[78] H. Benisty, D. Malah, and K. Crammer, "Non-parallel voice conversion using joint optimization of alignment by temporal context and spectral distortion," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 7909–7913.

[79] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning with application to clustering with side-information," in *Advances in Neural Information Processing System (NIPS)*, Vancouver, British Columbia, Canada, 2002, pp. 505–512.

[80] A. Mouchtaris, J. Van der Spiegel, and P. Mueller, "non-parallel training for voice conversion based on a parameter adaptation approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 952–963, 2006.

[81] T. Kinnunen, L. Juvela, P. Alku, and J. Yamagishi, “Non-parallel voice conversion using *i*-vector PLDA: Towards unifying speaker verification and transformation,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New Orleans, USA, 2017, pp. 5535–5539.

[82] L. Sun *et al.*, “Phonetic posteriorgrams for many-to-one voice conversion without parallel data training,” in *International Conference on Multimedia and Expo (ICME)*, Seattle, USA, 2016, pp. 1–6.

[83] K. Shikano, S. Nakamura, and M. Abe, “Speaker adaptation and voice conversion by codebook mapping,” in *IEEE International Sympoisum on Circuits and Systems*, Singapore, 1991, pp. 594–597.

[84] H. Matsumoto and Y. Yamashita, “Unsupervised speaker adaptation from short utterances based on a minimized fuzzy objective function,” *Journal of the Acoustical Society of Japan (E)*, vol. 14, no. 5, pp. 353–361, 1993.

[85] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[86] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, “Voice conversion using partial least squares regression,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[87] 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 *International Conference on Spoken Language Processing (ICSLP)*, Colorado, USA, 2002, pp. 285–288.

[88] R. H. Laskar, F. A. Talukdar, R. Bhattacharjee, and S. Das, “Voice conversion by mapping the spectral and prosodic features using support vector machine,” in *Applications of Soft Computing*. Springer, 2009, pp. 519–528.

[89] D. Sundermann, H. Ney, and H. Hoge, “VTLN-based cross-language voice conversion,” in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, St. Thomas, U.S, 2003, pp. 676–681.

[90] D. Erro, E. Navas, and I. Hernáez, “Iterative MMSE estimation of vocal tract length normalization factors for voice transformation,” in *INTERSPEECH*, Oregon, USA, 2012, pp. 86–89.

[91] T. Toda, H. Saruwatari, and K. Shikano, “Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, UT, USA, 2001, pp. 841–844.

[92] E. Godoy, O. Rosec, and T. Chonavel, “Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or non-parallel corpora,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1313–1323, 2012.

[93] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li, “Exemplar-based voice conversion using non-negative spectrogram deconvolution,” in *ISCA Speech Synthesis Workshop (SSW)*, Barcelona, Spain, 2013, pp. 201–206.

[94] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, “Exemplar-based sparse representation with residual compensation for voice conversion,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1506–1521, 2014.

[95] R. Takashima, T. Takiguchi, and Y. Ariki, “Exemplar-based voice conversion in noisy environment,” in *IEEE Spoken Language Technology (SLT) Workshop*, Miami, Florida, USA, 2012, pp. 313–317.

[96] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, “Voice conversion using artificial neural networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 3893–3896.

[97] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, “Spectral mapping using artificial neural networks for voice conversion,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 954–964, 2010.

[98] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning, 1^{st}, Eds.* The MIT Press, 2016.

[99] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, “Voice conversion using deep neural networks with layerwise generative training,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.

[100] T. Nakashika, T. Takiguchi, and Y. Ariki, “Voice conversion using RNN pre-trained by recurrent temporal restricted Boltzmann machines,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 23, no. 3, pp. 580–587, 2015.

[101] L. Sun, S. Kang, K. Li, and H. Meng, “Voice conversion using deep bidirectional long short-term memory based recurrent neural networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 4869–4873.

[102] J. Wu, D. Huang, L. Xie, and H. Li, “Denoising recurrent neural network for deep bidirectional LSTM based voice conversion,” in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3379–3383.

[103] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, “Generative adversarial network-based postfilter for statistical parametric speech synthesis,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Louisiana, USA, 2017, pp. 4910–4914.

[104] H. Miyoshi, Y. Saito, S. Takamichi, and H. Saruwatari, “Voice conversion using sequence-to-sequence learning of context posterior probabilities,” in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1268–1272.

[105] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, “Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 5274–5278.

[106] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, “Voice conversion from non-parallel corpora using variational autoencoder,” in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC)*, Jeju, Korea, 2016, pp. 1–6.

[107] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3364–3368.

[108] Y. Gao, R. Singh, and B. Raj, "Voice impersonation using generative adversarial networks," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 2506–2510.

[109] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba, "High quality non-parallel voice conversion based on cycle-consistent adversarial network," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 5279–5283.

[110] R. Kompe and R. Kompe, *Prosody in Speech Understanding Systems,* 1^{st} *Eds.* Springer, 1997, vol. 1307.

[111] Y. Xu, "Speech prosody: A methodological review," *Journal of Speech Sciences*, vol. 1, no. 1, pp. 85–115, 2011.

[112] D. T. Chappell and J. H. L. Hansen, "Speaker-specific pitch contour modeling and modification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seattle, WA, USA, 1998, pp. 885–888.

[113] N. J. Shah and H. A. Patil, "Novel filtering-based F_0 estimation algorithm with an application to voice conversion," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC)*. Kuala Lumpur, Malaysia: IEEE, 2017, pp. 1528–1531.

[114] Z.-Z. Wu, T. Kinnunen, E. S. Chng, and H. Li, "Text-independent f_0 transformation with non-parallel data for voice conversion," in *INTERSPEECH*, Makuhari, Chiba, Japan, 2010, pp. 1732–1735.

[115] K. S. Rao, "Voice conversion by mapping the speaker-specific features using pitch synchronous approach," *Computer Speech & Language*, vol. 24, no. 3, pp. 474–494, 2010.

[116] K. S. Rao and B. Yegnanarayana, "Voice conversion by prosody and vocal tract modification," in *International Conference on Information Technology (ICIT)*, Bhubaneshwar, India, 2006, pp. 111–116.

[117] B. Şişman, H. Li, and K. C. Tan, "Transformation of prosody in voice conversion," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 1537–1546.

[118] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The Voice Conversion Challenge 2016," in *INTERSPEECH*, San Fransisco, USA, 2016, pp. 1–5.

[119] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and non-parallel methods," in *The Speaker and Language Recognition Workshop Odyssey*, Les Sables d'Olonne, France, 2018, pp. 195–202.

[120] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, “Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools,” *IEEE signal processing magazine*, vol. 32, no. 2, pp. 114–124, 2015.

[121] H. B. Sailor, “Objective evaluation of speech quality of text-to-speech (TTS) synthesis,” M.Tech Thesis, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, India, 2013.

[122] T. Kinnunen, J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, and Z. Ling, “A spoofing benchmark for the 2018 voice conversion challenge: Leveraging from spoofing countermeasures for speech artifact assessment,” in *The Speaker and Language Recognition Workshop Odyssey*, Les Sables d‘Olonne, France, 2018, pp. 187–194.

[123] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[124] Z. Jin, A. Finkelstein, S. DiVerdi, J. Lu, and G. J. Mysore, “Cute: A concatenative method for voice conversion using exemplar-based unit selection,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5660–5664.

[125] H. Ming, D. Huang, L. Xie, S. Zhang, M. Dong, and H. Li, “Exemplar-based sparse representation of timbre and prosody for voice conversion,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5175–5179.

[126] N. Maeda, H. Banno, S. Kajita, K. Takeda, and F. Itakura, “Speaker conversion through nonlinear frequency warping of straight spectrum,” in *EUROSPEECH*, Budapest, Hungary, 1999, pp. 1–4.

[127] D. Erro, A. Moreno, and A. Bonafonte, “Voice conversion based on weighted frequency warping,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2010.

[128] X. Tian, Z. Wu, S. W. Lee, and E. S. Chng, “Correlation-based frequency warping for voice conversion,” in *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Singapore, 2014, pp. 211–215.

[129] Y. Agiomyrgiannakis and Z. Roupakia, “Voice morphing that improves TTS quality using an optimal dynamic frequency warping-and-weighting transform,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, 2016, pp. 5650–5654.

[130] T. C. Zorilă, D. Erro, and I. Hernaez, “Improving the quality of standard GMM-based voice conversion systems by considering physically motivated linear transformations,” in *Advances in Speech and Language Technologies for Iberian Languages, Springer*, 2012, pp. 30–39.

[131] X. Tian, Z. Wu, S. W. Lee, N. Q. Hy, M. Dong, and E. S. Chng, “System fusion for high performance voice conversion,” in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2759–2763.

[132] D. Mostefa, O. Hamon, N. Moreau, and K. Choukri, "Evaluation Report Deliverable D30 of the EU funded projects TC-STAR, 2007," http://tcstar.org/pages/ThirdEvaluationCampaign.htm, {Last Accessed: May 05, 2019 }.

[133] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International Conference on Artificial Intelligence and Statistics*, Sardinia, Italy, 2010, pp. 249–256.

[134] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *Journal of Machine Learning Research*, vol. 11, no. Feb, pp. 625–660, 2010.

[135] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[136] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[137] G. Cheng, V. Peddinti, D. Povey, V. Manohar, S. Khudanpur, and Y. Yan, "An exploration of dropout with LSTMs," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1586–1590.

[138] H. B. Sailor and H. A. Patil, "Auditory feature representation using convolutional restricted Boltzmann machine and Teager energy operator for speech recognition," *JASA-EL*, vol. 141, no. 6, pp. 500–506, 2017.

[139] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, Ft. Lauderdale, FL, USA, 2011, pp. 315–323.

[140] D. Kingma and J. Ba, "ADAM: A method for stochastic optimization," in *International Confernece on Learning Representation (ICLR)*, San Diego, USA, 2015, pp. 1–15.

[141] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of ADAM and beyond," in *International Conference on Learning Representations (ICLR)*, Vancouver, CANADA, 2018, pp. 1–23.

[142] N. Shah, H. A. Patil, and M. H. Soni, "Time-frequency mask-based speech enhancement using convolutional generative adversarial network," in *Asia-Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference*, Honolulu, Hawaii, USA, 2018, pp. 1246–1251.

[143] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3642–3646.

[144] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1283–1287.

[145] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 26, no. 1, pp. 84–96, 2018.

[146] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *International Conference on Machine Learning (ICML)*, Sydney, Australia, 2017, pp. 1857–1865.

[147] C. C. Aggarwal, "Outlier analysis," in *Data Mining*. Springer, 2015, pp. 237–263.

[148] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[149] P. Rousseeuw and A. Leroy, *Robust Regression and Outlier Detection,*1^{st}*, Eds.* John Wiley& Sons, New York, 1987.

[150] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, 2004.

[151] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, pp. 2250–2267, 2014.

[152] A. Richter, "Modeling of continuous speech observations," in *Advances in Speech Processing Conference*, IBM Europe Institute in Oberlech, Austria, 1986, pp. 1–4.

[153] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the STFT-phase," *IEEE Signal Processing Letters*, vol. 20, no. 2, pp. 129–132, 2013.

[154] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*. Springer, 2005, vol. 2.

[155] P. Mahalanobis, "Mahalanobis distance," *Proceedings National Institute of Science of India*, vol. 49, no. 2, pp. 234–256, 1936.

[156] P. J. Rousseeuw and M. Hubert, "Robust statistics for outlier detection," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 73–79, 2011.

[157] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.

[158] M. Hubert, P. J. Rousseeuw, and K. Vanden Branden, "ROBPCA: A new approach to robust principal component analysis," *Technometrics*, vol. 47, no. 1, pp. 64–79, 2005.

[159] M. Hubert, P. Rousseeuw, and T. Verdonck, "Robust PCA for skewed data and its outlier map," *Computational Statistics & Data Analysis*, vol. 53, no. 6, pp. 2264–2274, 2009.

[160] M. Hubert and M. Debruyne, "Minimum covariance determinant," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 1, pp. 36–43, 2010.

[161] J. Hardin and D. M. Rocke, "The distribution of robust distances," *Journal of Computational and Graphical Statistics*, vol. 14, no. 4, pp. 928–946, 2005.

[162] A. R. Webb, *Statistical Pattern Recognition*, 1^{st}*, Eds.* John Wiley & Sons, 2003.

[163] J. J. Wolf, “Efficient acoustic parameters for speaker recognition,” *The J. of the Acoust. Soc. of Amer. (JASA)*, vol. 51, no. 6B, pp. 2044–2056, 1972.

[164] J. P. Campbell, “Speaker recognition: A tutorial,” *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[165] X. Lu and J. Dang, “An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification,” *Speech Communication*, vol. 50, no. 4, pp. 312–322, 2008.

[166] K. Sriskandaraja, V. Sethu, E. Ambikairajah, and H. Li, “Front-end for antispoofing countermeasures in speaker verification: Scattering spectral decomposition,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 632–643, 2017.

[167] L. Li, Y. Chen, D. Wang, and T. F. Zheng, “A study on replay attack and anti-spoofing for automatic speaker verification,” in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 92–96.

[168] Q. Lin, E.-E. Jan, C. Che, D.-S. Yuk, and J. Flanagan, “Selective use of the speech spectrum and a VQGMM method for speaker identification,” in *Fourth International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, USA, 1996, pp. 2415–2418.

[169] J. E. Freund and I. Miller, *John E. Freund’s Mathematical Statistics: With Applications*, 1^{st}*, Eds.* Pearson Education India, 2004.

[170] “Few Converted Samples,” URL:https://drive.google.com/open?id=1U-zMhdTY6XeRnIbvc1EzdFGYJc1CKlUv, {Last Accessed: May 18, 2019}.

[171] I. Rec, “P. 85. A method for subjective performance assessment of the quality of speech voice output devices,” *International Telecommunication Union (ITU), Geneva*, Available Online:{https://www.itu.int/rec/T-REC-P.85-199406-I/en} Last Accessed{May 26, 2019}.

[172] N. J. Shah and H. A. Patil, “Analysis of features and metrics for alignment in text-dependent voice conversion,” in *International Conference on Pattern Recognition and Machine Intelligence (PReMI), ISI, Kolkata: B. Uma Shankar et. al., Lecture Notes in Computer Science (LNCS), Springer*, vol. 10597, 2017, pp. 299–307.

[173] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, “COVAREP: A collaborative voice analysis repository for speech technologies,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 960–964.

[174] T. Drugman, Y. Stylianou, Y. Kida, and M. Akamine, “Voice activity detection: Merging source and filter-based information,” *IEEE Signal Processing Letters*, vol. 23, no. 2, pp. 252–256, 2016.

[175] P. Zhan and A. Waibel, “Vocal tract length normalization for large vocabulary continuous speech recognition,” CMU Computer Science Technical Report, Pittsburgh, USA, Tech. Rep., 1997.

[176] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 930–944, 2005.

[177] R. G. Bartle and D. R. Sherbert, *Introduction to Real Analysis, 1^{st}, Eds.* Wiley New York, 1992.

[178] E. Kreyszig, *Introductory Functional Analysis with Applications, 1^{st}, Eds.* wiley New York, 1989, vol. 81.

[179] H. A. Patil and T. K. Basu, "Detection of bilingual twins by Teager energy based features," in *International Conference on Signal Processing and Communications (SPCOM)*, Bangalore, India, 2004, pp. 32–36.

[180] S. Furui, "On the role of spectral transition for speech perception," *The Journal of the Acoust. Soc. of Amer. (JASA)*, vol. 80, no. 4, pp. 1016–1025, 1986.

[181] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoust., Speech, and Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.

[182] K. P. Murphy, *Machine Learning : A Probabilistic Perspective.* The MIT Press, first, 2013.

[183] P. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Patern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[184] A. Gunawardana and W. Byrne, "Convergence theorems for generalized alternating minimization procedures," *Journal of Machine Learning Research*, vol. 6, no. 1, pp. 2049–2073, 2005.

[185] H. A. Patil, P. Dutta, and T. Basu, "On the investigation of spectral resolution problem for identification of female speakers in bengali," in *IEEE International Conference on Industrial Technology (ICIT)*, Mumbai,India, 2006, pp. 375–380.

[186] A. E. Rosenberg, "Automatic speaker verification: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 475–487, 1976.

[187] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," *arXiv preprint arXiv:1611.02200*, 2016.

[188] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.

[189] B. Kulis *et al.*, "Metric learning: A survey," *Foundations and Trends® in Machine Learning*, vol. 5, no. 4, pp. 287–364, 2013.

[190] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning with application to clustering with side-information," in *Neural Information Processing Systems (NIPS)*, Vancouver, British Columbia, Canada, 2002, pp. 505–512.

[191] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Advances in Neural Information Processing Systems*, Vancouver, British Columbia, Canada, 2005, pp. 1473–1480.

[192] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic speech database," *National Institute of Standards and Technology (NIST),USA*, vol. 15, pp. 29–50, 1988.

[193] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory,*1^{st}*, Eds.* Upper Saddle River, New Jersey: Prentic Hall, 1998.

[194] H. A. Patil, T. Patel, S. Talesara, N. Shah, H. Sailor, B. Vachhani, J. Akhani, B. Kanakiya, Y. Gaur, and V. Prajapati, "Algorithms for speech segmentation at syllable-level for text-to-speech synthesis system in Gujarati," in *Oriental COCOSDA*, New Delhi, India, 2013, pp. 1–7.

[195] B. B. Vachhani and H. A. Patil, "Use of PLP cepstral features for phonetic segmentation," in *International Conference on Asian Language Processing (IALP)*, 2013, pp. 143–146.

[196] B. B. Vachhani, C. Bhat, and S. Kopparapu, "Robust phonetic segmentation using multi-taper spectral estimation for noisy and clipped speech," in *European Signal Processing Conference (EUSIPCO)*, Budapest, Hungary, 2016, pp. 1343–1347.

[197] B. B. Vachhani, C. Bhat, and S. Kopparapu, *Phonetic segmentation using knowledge from visual and perceptual domain.* Prague, Czech Republic: Ekštein et. al. (Eds), Lecture Notes in Computer Science (LNCS), Springer, Text, Speech, and Dialogue (TSD), vol. 10415, pp. 393–401, 2017.

[198] M. C. Madhavi, H. A. Patil, and B. B. Vachhani, "Spectral transition measure for detection of obstruents," in *European Signal Processing Conference (EUSIPCO)*, Nice, France, 2015, pp. 330–334.

[199] M. K. Leonard and E. F. Chang, "Dynamic speech representations in the human temporal lobe," *Trends in Cognitive Sciences*, vol. 18, no. 9, pp. 472–479, 2014.

[200] B. Khalighinejad, G. C. da Silva, and N. Mesgarani, "Dynamic encoding of acoustic features in neural responses to continuous speech," *Journal of Neuroscience*, vol. 37, no. 8, pp. 2176–2185, 2017.

[201] B. Delgutte, "Auditory neural processing of speech," *The Handbook of Phonetic Sciences*, pp. 507–538, 1997.

[202] E. Seifritz, F. Esposito, F. Hennel, H. Mustovic *et al.*, "Spatiotemporal pattern of neural processing in the human auditory cortex," *Science*, vol. 297, no. 5587, p. 1706, 2002.

[203] A. Moller, *Auditory Physiology,* 1^{st}*, Eds.* Elsevier, 2012.

[204] S. Dusan and L. R. Rabiner, "On the relation between maximum spectral transition positions and phone boundaries," in *INTERSPEECH*, Pittsburgh, USA, 2006, pp. 17–21.

[205] N. J. Shah and H. A. Patil, "On the convergence of INCA algorithm," in *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC)*, Kuala Lumpur, Malaysia, 2017, pp. 559–562.

[206] E. Godoy, O. Rosec, and T. Chonavel, "Alleviating the one-to-many mapping problem in voice conversion with context-dependent modelling," in *INTERSPEECH*, Brighton, United Kingdom, 2009, pp. 1627–1630.

[207] S. H. Mohammadi, "Reducing one-to-many problem in voice conversion by equalizing the formant locations using dynamic frequency warping," *arXiv preprint arXiv:1510.04205*, 2015.

[208] O. Turk and L. M. Arslan, "Robust processing techniques for voice conversion," *Computer Speech & Language, Elsevier*, vol. 20, no. 4, pp. 441–467, 2006.

[209] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Non-parallel training for many-to-many eigenvoice conversion," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010, pp. 4822–4825.

[210] C. H. Lee and C. H. Wu, "MAP-based adaptation for speech conversion using adaptation data selection and non-parallel training," in *INTERSPEECH*, Pittsburgh, USA, 2006, pp. 2254–2257.

[211] J. Pinto and H. Hermansky, "Combining evidence from a generative and a discriminative model in phoneme recognition," in *INTERSPEECH*, Brisbane, Australia, 2008.

[212] G. Aradilla, J. Vepa, and H. Bourlard, "Using posterior-based features in template matching for speech recognition," in *INTERSPEECH*, Pittsburgh, PA, USA, 2006, pp. 2570–2573.

[213] J. Pinto, G. S. V. S. Sivaram, M. Magimai-Doss, H. Hermansky, and H. Bourlard, "Analysis of MLP-based hierarchical phoneme posterior probability estimator," *IEEE Transactions Audio, Speech & Language Processing*, vol. 19, no. 2, pp. 225–241, 2011.

[214] S. Thomas, S. Ganapathy, A. Jansen, and H. Hermansky, "Data-driven posterior features for low resource speech recognition applications," in *INTERSPEECH*, Portland, Oregon, USA, 2012, pp. 791–794.

[215] P. Schwarz, P. Matějka, and J. Černocký, *Towards lower error rates in phoneme recognition*. Berlin, Heidelberg: Springer, 2004, pp. 465–472.

[216] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, Merano, Italy, 2009, pp. 398–403.

[217] A. Mandal, K. P. Kumar, and P. Mitra, "Recent developments in spoken term detection: A survey," *International Journal of Speech Technology (IJST), Springer*, vol. 17, no. 2, pp. 183–198, 2014.

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

[219] Y. Kim and J. Smith, "A speech feature based on Bark frequency warping-the non-uniform linear prediction (NLP) cepstrum," in *IEEE Workshop on Applications of Signal Process. to Audio & Acoust*, New Paltz, NY, 1999, pp. 131–134.

[220] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *the J. of the Acoust. Soc. of Amer. (JASA)*, vol. 87, no. 4, pp. 1738–1752, 1990.

[221] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning (ICML)*, Lille, France, 2015, pp. 448–456.

[222] I. Danihelka, B. Lakshminarayanan, B. Uria, D. Wierstra, and P. Dayan, “Comparison of maximum likelihood and GAN-based training of real NVPs,” *arXiv preprint arXiv:1705.05263*, 2017, {Last Accessed: March 15, 2018}.

[223] H. B. Sailor and H. A. Patil, “Fusion of magnitude and phase-based features for objective evaluation of TTS voice,” in *IEEE International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Singapore, 2014, pp. 521–525.

[224] T. Ganchev, A. Lazaridis, I. Mporas, and N. Fakotakis, “Performance evaluation for voice conversion systems,” in *International Conference on Text, Speech and Dialogue*. Berlin, Germany: Springer, 2008, pp. 317–324.

[225] S. Aryal and R. Gutierrez-Osuna, “Articulatory inversion and synthesis: Towards articulatory-based modification of speech,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 7952–7956.

[226] A. W. Black, H. T. Bunnell, Y. Dou, P. K. Muthukumar, F. Metze, D. Perry, T. Polzehl, K. Prahallad, S. Steidl, and C. Vaughn, “Articulatory features for expressive speech synthesis,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Japan, 2012, pp. 4005–4008.

[227] D. Childers, B. Yegnanarayana, and K. Wu, “Voice conversion: Factors responsible for quality,” in *International Conference on Acoustics, Speech, and Signal Processing*, Florida, USA, 1985, pp. 748–751.

[228] J. Wang, J. R. Green, and A. Samal, “Individual articulator‘s contribution to phoneme production,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, BC, Canada, 2013, pp. 7785–7789.

[229] P. J. Jackson and V. D. Singampalli, “Statistical identification of articulation constraints in the production of speech,” *Speech Communication*, vol. 51, no. 8, pp. 695–710, 2009.

[230] P. Ladefoged and K. Johnson, *A Course in Phonetics, 6th, Eds.* Nelson Education, 2014.

[231] R. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, Victoria, British Columbia, Canada, 1993, pp. 125–128.

[232] P. Lanchantin and X. Rodet, “Objective evaluation of the dynamic model selection method for spectral voice conversion,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Czech Republic, 2011, pp. 5132–5135.

[233] K. Richmond, “Estimating articulatory parameters from the acoustic speech signal,” Ph.D. Thesis, The University of Edinburgh, 2002.

[234] M. Li, J. Kim, P. K. Ghosh, V. Ramanarayanan, and S. Narayanan, “Speaker verification based on fusion of acoustic and articulatory information,” in *INTERSPEECH*, Lyon, France, 2013, pp. 1614–1618.

[235] P. K. Ghosh and S. Narayanan, “Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion,” *The Journal of the Acoustical Society of America (JASA)*, vol. 130, no. 4, pp. EL251–EL257, 2011.

[236] S. Aryal and R. Gutierrez-Osuna, "Articulatory-based conversion of foreign accents with deep neural networks," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 3385–3389.

[237] A. R. Toth and A. W. Black, "Using articulatory position data in voice transformation," in *ISCA Speech Synthesis Workshop (SSW6)*, Bonn , Germany, 2007, pp. 182–187.

[238] A. A. Wrench, "A multichannel articulatory database and its application for automatic speech recognition," in *Seminar of Speech Production (SSP)*, Kloster Seeon, Germany, 2000, pp. 305–308.

[239] A. Rajpal and H. A. Patil, "Jerk minimization for acoustic-to-articulatory inversion," in *Speech Synthesis Workshop (SSW)*, USA, 2016, pp. 82–87.

[240] P. K. Ghosh and S. Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America (JASA)*, vol. 128, no. 4, pp. 2162–2172, 2010.

[241] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. First (Eds.), John Wiley & Sons, 2012.

[242] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066138, 2004.

[243] V. C. Tartter, "What is in a whisper," *The J. of the Acoust. Soc. of Amer. (JASA)*, vol. 86, no. 5, pp. 1678–1683, 1989.

[244] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, 2003, pp. 704–708.

[245] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 10, pp. 2448–2458, 2010.

[246] S. Ghaffarzadegan, H. Bořil, and J. H. L. Hansen, "Generative modeling of pseudo-whisper for robust whispered speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1705–1720, 2016.

[247] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Communication*, vol. 52, no. 4, pp. 288–300, 2010.

[248] T. Toda and K. Shikano, "NAM-to-speech conversion with Gaussian mixture models," in *INTERSPEECH*, Lisbon, Portugal, 2005, pp. 1957–1960.

[249] V.-A. Tran, G. Bailly, H. Lœvenbruck, and T. Toda, "Multimodal HMM-based NAM-to-speech conversion," in *INTERSPEECH*, Brighton, United Kingdom (UK), 2009, pp. 656–659.

[250] V. A. Tran, G. Bailly, H. Lœvenbruck, and T. Toda, "Improvement to a NAM-captured whisper-to-speech system," *Speech Communication*, vol. 52, no. 4, pp. 314–326, 2010.

[251] H. Konno, M. Kudo, H. Imai, and M. Sugimoto, "Whisper-to-normal speech conversion using pitch estimated from spectrum," *Speech Communication*, vol. 83, pp. 10–20, 2016.

[252] M. Janke, M. Wand, T. Heistermann, T. Schultz, and K. Prahallad, "Fundamental frequency generation for whisper-to-audible speech conversion," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 2579–2583.

[253] G. N. Meenakshi and P. K. Ghosh, "Whispered speech to neutral speech conversion using bidirectional LSTMs," in *INTERSPEECH*, Hyderabad, India, 2018, pp. 491–495.

[254] W. Meyer-Eppler, "Realization of prosodic features in whispered speech," *The J. of the Acoust. Soc. of Amer. (JASA)*, vol. 29, no. 1, pp. 104–106, 1957.

[255] T. Toda, K. Nakamura, T. Nagai, T. Kaino, Y. Nakajima, and K. Shikano, "Technologies for processing body-conducted speech detected with non-audible murmur microphone," in *INTERSPEECH*, Brighton, United Kingdom, 2009, pp. 632–635.

[256] M. Airaksinen, T. Bäckström, and P. Alku, "Automatic estimation of the lip radiation effect in glottal inverse filtering," in *INTERSPEECH*, Singapore, 2014, pp. 398–402.

[257] "Publicly available: The CSTR NAM TIMIT Plus Corpus," URL: homepages.inf.ed.ac.uk/jyamagis/release/CSTR-NAM-TIMIT-Plus-ver0.81.tar.gz, {Last Accessed: May 18, 2019}.

[258] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1097–1111, 2008.

[259] M. Wand, M. Janke, and T. Schultz, "The EMG-UKA corpus for electromyographic speech processing," in *INTERSPEECH*, Singapore, 2014, pp. 1–5.

[260] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 16, no. 1, pp. 229–238, 2008.

[261] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach, 1st, Eds.* Springer Science & Business Media, 2012.

[262] N. J. Shah, M. Zaki, and H. A. Patil, "Influence of various asymmetrical contextual factors for TTS in a low resource language," in *International Conference on Asian Language Processing (IALP)*, Singapore, 2014, pp. 107–110.

[263] M. H. Davis, M. A. Ford, F. Kherif, and I. S. Johnsrude, "Does semantic context benefit speech understanding through "top–down" processes? Evidence from time-resolved sparse fMRI," *Journal of Cognitive Neuroscience*, vol. 23, no. 12, pp. 3914–3932, 2011.

[264] L. L. Holt and A. J. Lotto, "Speech perception within an auditory cognitive science framework," *Current Directions in Psychological Science*, vol. 17, no. 1, pp. 42–46, 2008.

[265] M. Chait, D. Poeppel, A. De Cheveigné, and J. Z. Simon, "Processing asymmetry of transitions between order and disorder in human auditory cortex," *Journal of Neuroscience*, vol. 27, no. 19, pp. 5207–5214, 2007.

[266] S. Yang, L. Xie, X. Chen, X. Lou, X. Zhu, D. Huang, and H. Li, "Statistical parametric speech synthesis using generative adversarial networks under a multi-task learning framework," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Okinawa, Japan, Dec 2017, pp. 685–691.

[267] B. Sisman, M. Zhang, and H. Li, "Group sparse representation with wavenet vocoder adaptation for spectrum and prosody conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1085–1097, 2019.

[268] Z. Liu, Y. Huang, and J. Huang, "Patchwork-based audio watermarking robust against de-synchronization and recapturing attacks," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1171–1180, 2019.

[269] J. W. Schnupp, I. Nelken, and A. J. King, *Auditory Neuroscience: Making Sense of Sound, 1st, Eds.* The MIT Press, 2012.

List of Publications from Thesis

Journal Papers

1. Nirmesh J. Shah and Hemant A. Patil, “Novel outliers removal approach for parallel voice conversion”, in Computer Speech & Language, Elsevier, vol. 58, November, pp. 127-152, 2019.
2. Nirmesh J. Shah, Mihir Parmar, Neil Shah and Hemant A. Patil, “Effectiveness of GANs in cross-domain non-audible murmur to speech conversion", *article under preparation*.

Book Chapters

3. Nirmesh J. Shah and Hemant A. Patil, “ Non-audible murmur to audible speech conversion," in Voice Technologies for Reconstruction and Enhancement, H. A. Patil and A. Neustein, Eds. De Gruyter Series in Speech Technology and Text Analytics in Medicine and Healthcare, 2020, pp. 125-150.
4. Nirmesh J. Shah and Hemant A. Patil,“Analysis of features and metrics for alignment in text-dependent voice conversion” in B. Uma Shankar et. al. (Eds), Lecture Notes in Computer Science (LNCS), Springer, PReMI, vol. 10597, pp. 299–307, 2017.

Conference Papers

5. Nirmesh J. Shah, Hardik B. Sailor and Hemant A. Patil, “Whether to pretrain DNN or Not?: An empirical analysis for voice conversion", in INTERSPEECH, Graz, Austria, 2019, pp. 639-643.
6. Nirmesh J. Shah and Hemant A. Patil, “Phone aware nearest neighbor technique using spectral transition measure for non-parallel voice conversion", in INTERSPEECH, Graz, Austria, 2019, pp. 1586-1590.
7. Nirmesh J. Shah and Hemant A. Patil,“Novel metric learning for non-parallel voice conversion", in *International Conference on Acoustics, Speech and Signal Proceesing (ICASSP)*, Brighton, UK, 2019, pp. 3722-3726.
8. Maitreya Patel, Mihir Parmar, Savan Doshi, Nirmesh Shah, and Hemant A. Patil, “Adaptive generative adversarial network for voice conversion," accepted in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC), Lanzhou, China, September 2019.

9. Mihir Parmar, Savan Doshi, Nirmesh Shah, Maitreya Patel and Hemant A. Patil, “Effectiveness of cross-domain architectures for whispered-to-normal speech conversion," in European Signal Processing Conference (EUSIPCO), A Coruna, Spain, 2019, pp. 1-5.

10. Nirmesh J. Shah and Hemant A. Patil, "Effectiveness of dynamic features in INCA and temporal context-INCA", in INTERSPEECH, Hyderabad, India, 2018, pp. 711–715.

11. Nirmesh J. Shah, Maulik C. Madhavi and Hemant A. Patil, “Unsupervised vocal tract length warped posterior features for non-parallel voice conversion", in INTERSPEECH, Hyderabad, India, 2018, pp. 1968-1972.

12. Neil Shah, Nirmesh J. Shah, and Hemant A. Patil, "Effectiveness of generative adversarial network for non-audible murmur-to-whisper speech conversion", in INTERSPEECH, Hyderabad, India, 2018, pp. 3157–3161.

13. Nirmesh J. Shah, Mihir Parmar, Neil Shah and Hemant A. Patil, "Novel MMSE DiscoGAN for cross-domain whisper-to-speech conversion", in Machine Learning in Speech and Language Processing (MLSLP) Workshop, Google Office, Hyderabad, 2018, pp. 1–3.

14. Nirmesh J. Shah, Sreeraj R., Neil Shah, Hemant A. Patil, “Novel unsupervised inter mixture weighted GMM posteriorgram for DNN and GAN-based voice conversion framework”, in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC), Honolulu, Hawaii, 2018, pp. 1776–1781.

15. Nirmesh J. Shah, Pramod B. Bachhav and Hemant A. Patil,“A novel filtering-based F_0 estimation algorithm with an application to voice conversion", in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC), Kuala Lumpur, Malaysia, 2017, pp. 1528-1531.

16. Nirmesh J. Shah and Hemant A. Patil,“On the convergence of INCA algorithm", in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC), Kuala Lumpur, Malaysia, 2017, pp. 559-562.

17. Nirmesh J. Shah and Hemant A. Patil,“Novel amplitude scaling method for bilinear frequency warping-based voice conversion", in IEEE International Conference on Acoustic, Speech and Signal Proceesing (ICASSP), New Orleans, 2017, pp. 5520-5524.

18. Avni Rajpal, Nirmesh J. Shah, Mohammadi Zaki and Hemant A. Patil,“Quality assessment of voice converted speech using articulatory features", in International

Conference on Acoustic, Speech, and Signal Proceesing (ICASSP), New Orleans, 2017,pp. 5515-5519.

19. Sushant V. Rao, Nirmesh J. Shah and Hemant A. Patil, “Novel pre-processing using outlier removal in voice conversion", in 9^{th} ISCA Speech Synthesis Workshop (SSW), Sunnyvale, CA, USA, 2016, pp.147-152.

20. Nirmesh J Shah, Bhavik Vachhani, Hardik Sailor and Hemant A. Patil, “Effectiveness of PLP-based phonetic segmentation algorithms for speech synthesis”, in Proc. of International Conference on Acoustic, Speech and Signal Proceesing (ICASSP), 2014, Florence, pp. 270-274.

Submitted Papers

21. Maitreya Patel, Mirali Purohit, Mihir Parmar, Nirmesh Shah and Hemant A. Patil, “AdaGAN: Adaptive GAN for Many-to-Many Non-Parallel Voice Conversion“, submitted for possible publications in International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26-30 April, 2020.

22. Nirmesh Shah, Savan Doshi, Mihir Parmar, Satyam Kumar, and Hemant A. Patil, “Do we really need alignment step in non-parallel voice conversion?," *submitted for possible publication* in ICASSP, Barcelona, Spain, May 2020.

23. Neil Shah, Sreeraj R., Nirmesh J. Shah and Hemant A. Patil, “Query-by-example spoken term detection using generative adversarial network," *submitted for possible publication* in ICASSP, Barcelona, Spain, May 2020.

24. Nirmesh J. Shah, M. Ali Basha Shaik, Periyasamy P., Hemant A. Patil and Vikram Vij, “Exploiting phase-based features for early robust detection of whispered speech”, *submitted for possible publication* in ICASSP, Barcelona, Spain, May 04-08, 2020.

25. Mihir Parmar, Maitreya Patel, Savan Doshi, Nirmesh Shah, Jui Shah, Mirali Purohit, and Hemant A. Patil, “Non-parallel whispered-to-normal speech conversion," *rejected* in INTERSPEECH, Graz, Austria, September 2019.

26. Savan Doshi, Mihir Parmar, Nishi Mehta, Nirmesh Shah, and Hemant A. Patil, “Non-parallel many-to-many voice conversion using augmented CycleGAN," *rejected* in INTERSPEECH, Graz, Austria, September 2019.

27. Nirmesh J. Shah, Neil Shah, Maulik C. Madhavi, Madhu R. Kamble, Hardik B. Sailor, Meet H. Soni, Sreeraj R., Prasad A. Tapkir and Hemant A. Patil, “Unsupervised VTLN posterior features for parallel and non-parallel voice conversion," *rejected* in INTERSPEECH, Graz, Austria, September 2019.

Brief Biography

Nirmesh J. Shah received the B.E. degree from Govt. Engg. College (GEC), Surat, India, in 2010, and the M.Tech degree from Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, India in 2013. During January 2014 to May 2019, he did his doctoral studies at DA-IICT, Gandhinagar, India. He was also associated with the consortium project on Development of Text-to-Speech (TTS) Systems in Indian Languages-Phase-II from May 2012 - December 2015. He did research internship at Samsung R&D Institute, Bangalore, India. Recently, he joined Gaana.com as a data scientist. His research interests include voice conversion and speech synthesis.

He has published 25 research papers in top conferences and peer-reviewed journal. His research area includes Voice Conversion and Speech Synthesis. He is a student member of International Speech Communication Association (ISCA), and IEEE Signal Processing Society (SPS). He has served as a reviewer for IEEE/ACM Transactions on Audio, Speech, and Language Processing, IEEE Journal of Biomedical and Health Informatics, and IEEE Journal of Selected Topics in Signal Processing. He received IEEE SPS Travel Grant to present his research papers at ICASSP 2014, and ICASSP 2017, respectively. He also got ISCA travel grant to present his paper at MLSLP, 2018; a satellite event of INTERSPEECH 2018.

He also presented a research paper during Oriental COCOSDA 2013, New Delhi, India and INTERSPEECH 2018, Hyderabad, India. He was also very active student volunteer during organizations of consortium project review meeting, 02 CEP workshops, and ISCA supported Winter School and 05 summer schools all at DA-IICT, Gandhinagar, India. In addition, he also worked as a student volunteer for INTERSPEECH 2018.