SPECTRAL MAPPING USING MULTIVARIATE POLYNOMIAL MODELING FOR VOICE CONVERSION

THESIS

submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

by

Parveen Kumar Lehana

(Roll No. 00407304)

under the supervision of

Prof. P. C. Pandey

Department of Electrical Engineering

Indian Institute of Technology Bombay

January 2013

Indian Institute of Technology Bombay

Department of Electrical Engineering

Ph.D. Thesis Approval

The thesis entitled "**Spectral Mapping Using Multivariate Polynomial Modeling for Voice Conversion**" by **Parveen Kumar Lehana** is approved, after successful completion of *viva voce* examination, for the award of the degree of **Doctor of Philosophy**.

Supervisor: _____ (Prof. P. C. Pandey)

Internal Examiner: _____ (Prof. Preeti Rao)

External Examiner: _____ (Prof. Arun Kumar)

Chairman: _____ (Prof. P. P. Date)

Date: January 28, 2013

Place: Mumbai

DECLARATION

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Parveen Kumar Lehana

(00407304)

Date: **28th January 2012**

Dedicated to my mother

Parveen Kumar Lehana / Supervisor: Prof. P. C. Pandey, "Spectral mapping using multivariate polynomial modeling for voice conversion", *Ph.D.Thesis*, Department of Electrical Engineering, IIT Bombay, Jan. 2013.

ABSTRACT

Voice conversion involves modification of the speech of a source speaker to make it perceptually similar to that of a target speaker, using a mapping derived from speech material spoken by them. The thesis presents a technique for modifying spectral characteristics using a single mapping obtained by multivariate polynomial modeling of the relation between the acoustic spaces of the source and the target speakers. Each parameter for generating the target speech is modeled as a multivariate polynomial function of the parameters of the source speech. The set of these functions is obtained from the time aligned source and target feature vectors. Voice conversion of the source speech signal is carried out by applying the estimated mapping for modification of spectral characteristics along with pitch and time scaling. A pitch-synchronous implementation of harmonic plus noise model (HNM) is used as the analysis-synthesis platform for voice conversion.

Out of the various polynomial models investigated, multivariate quadratic model (MQM) was found to be most suited. The method was evaluated for voice conversion for four speaker pairs (male-male, female-female, male-female, and female-male) using parallel speech data for training. Removal of redundant feature vectors from the training data, using a distance threshold, was found to improve the estimation of transformation functions. Voice conversion using MQM and GMM (with 64 mixture components) using the same training and test data resulted in almost similar output quality and decrease in target-transformed distance, but MQM needed a much shorter computation time for estimation of the mapping. Subjective evaluation showed that combination of spectral modification and pitch scaling resulted in transformed speech having good quality and almost the same identity as the target. Averaged across listeners, the scores for identification of transformed speech as the target were 93% for same-gender conversion and 100% for cross-gender conversion. The proposed voice conversion system needs to be investigated using a larger number of speaker pairs, different types of speech material, and other speech analysis-synthesis platforms.

[blank]

CONTENTS

ABSTRACT i
CONTENTS iii
LIST OF FIGURES v
LIST OF TABLES ix
LIST OF SYMBOLS xi
LIST OF ABBREVIATIONS xiii

Chapters

1 INTRODUCTION 1

1.1 Problem overview 1
1.2 Research objective 2
1.3 Thesis outline 2

2 VOICE CONVERSION 5

2.1 Introduction 5
2.2 Vector quantization 6
2.3 Artificial neural networks 7
2.4 Gaussian mixture model 7
2.5 Hidden Markov model 11
2.6 Frequency warping 14
2.7 Speaker interpolation 16
2.8 Residual modification 17
2.9 Prosody modification 20
2.10 Summary 21

3 HARMONIC PLUS NOISE MODEL FOR SPEECH MODIFICATION 23

3.1 Introduction 23
3.2 Harmonic plus noise model 23
3.3 HNM implementation 27
3.4 Speech modifications using HNM 33
3.5 Investigations using HNM 41
3.6 Summary 51

4 SPECTRAL MAPPING USING MULTIVARIATE POLYNOMIAL MODELING 55

4.1 Introduction 55
4.2 Multivariate polynomial modeling 56
4.3 Estimation of transformation function 57
4.4 Transformation of the source speech 60

5 RESULTS AND DISCUSSION 63

5.1 Introduction 63
5.2 Speech material 63

5.3 Experiment I: Redundancy of feature vectors 64
5.4 Experiment II: Comparison of polynomial and GMM based transformations 65
5.5 Experiment III: Interpolation capabilities of MQM 67
5.6 Experiment IV: Subjective evaluation of MQM using MOS and XAB 71
5.7 Discussion 74

6 SUMMARY AND CONCLUSIONS 79

6.1 Introduction 79
6.2 Summary of the investigations 79
6.3 Conclusions 80
6.4 Suggestions for future work 81

Appendices

A RESULTS OF INVESTIGATIONS USING HNM 83

B GMM BASED SPECTRAL MAPPING 87

C EVALUATION METHODS 91

D INSTRUCTIONS FOR XAB-MOS TEST 99

REFERENCES 101

ACKNOWLEDGEMENTS 123

AUTHOR'S RESUME AND LIST OF PUBLICATIONS 125

CERTIFICATE OF COURSE WORK

LIST OF FIGURES

Figure 3.1 HNM based speech analysis.

Figure 3.2 HNM based speech synthesis.

Figure 3.3 HNM based speech modification.

Figure 3.4 HNM-3 based speech analysis.

Figure 3.5 HNM-3 based speech synthesis.

Figure 3.6 An example of pitch contour in pitch scaled speech signal (2:1).

Figure 3.7 An example of pitch contour in time scaled speech signal (2:1).

Figure 3.8 Investigation I: Effect of HNM variants. Spectrograms of the Hindi utterance (F_s = 16 kHz, duration = 7.37 s) /d̪ʰoːbiːn̪ dʒəb soːkər ʊt̪ʰt̪i t̪oː d̪ekʰt̪i ki tʃoːkaː saːpʰ pəd̪aː həi ɔːr bərt̪ən mənd̪ʒɛː hʊɛː həiːn̪/ spoken by a female speaker (F1). a) recorded, b) synth. using HNM-1, c) synth. using HNM-2, and d) synth. using HNM-3.

Figure 3.9 Investigation I: PESQ-MOS test scores of the synthesized utterances (with recorded as reference), averaged over the sets of utterances from male (M), female (F), and male and female (M&F) speakers, for HNM-1, HNM-2, and HNM-3 based analysis-synthesis.

Figure 3.10 Investigation II: Effect of F_m. Spectrograms of the segment /kər/(F_s = 16 kHz, duration = 0.230 s) of the Hindi utterance /d̪ʰoːbiːn̪ dʒəb soːkər ʊt̪ʰt̪i t̪oː d̪ekʰt̪i ki tʃoːkaː saːpʰ pəd̪aː həi ɔːr bərt̪ən mənd̪ʒɛː hʊɛː həiːn̪/ spoken by a female speaker (F1). a) recorded, b) synth. with original F_m, c) synth. with F_m = 4 kHz, and d) synth. with F_m = 1 kHz.

Figure 3.11 Investigation II: PESQ-MOS test scores of the synthesized utterances (with recorded as the reference), for synthesis using different values of F_m, averaged over male (M), female (F), M&F sets of utterances for HNM-3 based analysis-synthesis.

Figure 3.12 Investigation III: Effect of jitter in GCI estimation. Spectrograms of the segment /tʃoːkaː/(F_s = 16 kHz, duration = 0.439 s)in Hindi utterance /d̪ʰoːbiːn̪ dʒəb soːkər ʊt̪ʰt̪i t̪oː d̪ekʰt̪i ki tʃoːkaː saːpʰ pəd̪aː həi ɔːr bərt̪ən mənd̪ʒɛː hʊɛː həiːn̪/spoken by a male speaker (F1). a) recorded (1.58%), b) synth. with jitter = 2.92%, c) synth. with jitter = 6.10%, and d) synth. with jitter = 6.73%. The horizontal axis is normalized time and the vertical axis is normalized

frequency.

Figure 3.13 Investigation III: PESQ-MOS test scores for the synthesized utterances (with recorded as reference) for synthesis using different amount of jitter, averaged over male (M), female (F), M&F sets of utterances for HNM-3 based analysis-synthesis, b) scatter plot of jitter in 10 vowels of two male speakers, c) error magnitude difference of GCI locations obtained from speech and EGG signals for the cardinal vowel /a/.

Figure 3.14 Investigation III: Scatter plot of jitters obtained from speech and EGG signals in 10 vowels from two male speakers.

Figure 3.15 Investigation IV: Effect of input SNR on analysis-synthesis using HNM-3. Spectrograms of the Hindi utterance (F_s = 16kHz, duration = 7.37 s) /d̪ʰoːbiːn̪ dʒəb soːkər ʊt̪ʰt̪i t̪oː d̪ekʰt̪i ki tʃoːkaː saːpʰ pəd̪aː həi ɔːr bərt̪ən mən̪dʒɛː hʊɛː həiːn̪/ spoken by a female speaker (F1) with different SNR values. a) recorded, b) 18 dB, c) 6 dB, and d) 4 dB.

Figure 3.16 Investigation IV: Effect of input SNR on analysis-synthesis using HNM-3. PESQ-MOS test scores (with recorded speech as reference) averaged over male (M), female (F), M&F sets of utterances for HNM-3 based analysis-synthesis.

Figure 3.17 Investigation V: Effect of estimated phase. PESQ-MOS test scores for analysis-synthesis (AS), synthesized with estimated phase (EP), and synthesized with source phase (SP), averaged across sentences and four speakers. The standard deviation is shown by error bars. a) Male speaker set, b) Female speaker set.

Figure 3.18 Investigation VI: Pitch and time scaling. Spectrograms of the Hindi utterance (16 kHz, 7.37 s) /d̪ʰoːbiːn dʒəb soːkər ʊt̪ʰt̪i t̪oː d̪ekʰt̪i ki tʃoːkaː saːpʰ pəd̪aː həi ɔːr bərt̪ən mən̪dʒɛː hʊɛː həiːn̪/ spoken by a female speaker (F1). a) recorded, b) pitch scaled by a factor of 0.5, c) pitch scaled by a factor of 1.5, d) time scaled by a factor of 0.5, and e) time scaled by a factor of 1.5.

Figure 4.1 Estimation of transformation function using HNM.

Figure 4.2 Transformation of speech signal using HNM.

Figure 5.1 Exp. I: Number of feature vectors (voiced and unvoiced frames) as a function of distance used in grouping.

Figure 5.2 Exp. II: Cepstral Mahalanobis distance between the source/target (ST) and target-transformed/target (TT′) pairs. a) M1-M2, b) F1-F2, c) M3-F3, d) F4-M4.

Figure 5.3 Exp. III: Normalized average cepstral Mahalanobis distance between the target and transformed source for 54 speech segments using sets of transformation functions Fn0 to Fn54. The x-axis denotes the speech segment number.

Figure 5.4 Exp. IV: MOS test scores for the sets of source (S), target (T), pitch scaled (PS), spectral modification (SM), and spectral modification and pitch scaled (SMPS) utterances.

Figure 5.5 Exp. IV: XAB scores for the sets of source (S), target (T), pitch scaled (PS), spectral modification (SM), and spectral modification and pitch scaled (SMPS) utterances.

Figure C.1 Effect of white noise on the PESQ-MOS test scores on male (M) and female (F) speech signal.

[blank]

LIST OF TABLES

Table 5.1 Exp. II: Cepstral Mahalanobis distance between the source and target (ST) and the target and transformed (TT') for different speaker pairs (mean and standard deviation for six utterances).

Table 5.2 Exp. III: Speech segments used for training and testing.

Table 5.3 Exp. IV: MOS test scores for the sets of source (S), target (T), pitch scaled (PS), spectral modification (SM), and spectral modification and pitch scaled (SMPS) utterances.

Table 5.4 Exp. IV: XAB scores for the sets of source (S), target (T), pitch scaled (PS), spectral modification (SM), and spectral modification and pitch scaled (SMPS) utterances.

Table A.1 Investigation I: Effect of HNM variants. PESQ-MOS test scores, averaged over the utterances in the set, male (M), female (F), and M&F for HNM-1, HNM-2, and HNM-3 based analysis-synthesis (reference: recorded speech). SD: standard deviation.

Table A.2 Investigation II: Effect of maximum voiced frequency on synthesized speech using HNM-3 based analysis-synthesis: Mean and SD of PESQ-MOS test scores for synthesized utterances using different values of F_m (reference: recorded speech), averaged over male (M), female (F), M&F sets of utterances.

Table A.3 Investigation III: Effect of jitter introduced in GCI estimation: PESQ-MOS test scores for synthesis using different amount of jitter, averaged over male (M), female (F), M&F sets of utterances for HNM-3 based analysis-synthesis (reference: recorded speech). γ = jitter control factor.

Table A.4 Investigation IV: Effect of input SNR on HNM-3 based analysis-synthesis: PESQ-MOS test scores averaged over male (M), female (F), M&F sets of utterances (reference: recorded speech).

Table A.5 Investigation V: Effect of phase estimation methods. PESQ-MOS test scores averaged over 24 utterances (6 utterances $\times$ 4 speaker pairs) for male and female speech (reference: recorded speech).

[blank]

LIST OF SYMBOLS

$\mathbf{a}$	column vector of complex harmonic amplitudes
$a_{l,i}(n)$	interpolated magnitude for frame i and harmonic l
$a_l(n)$	magnitude of harmonic l at sample n
$b_l(n)$	time varying complex amplitudes
$\mathbf{c}$	column vector of the cepstral coefficients
$c_r(n)$	cepstral coefficient for reference speaker
$c_t(n)$	cepstral coefficient for test speaker
c_m	cepstral coefficient m
$d(t,r)$	distance measure between test and reference signal
$D(t,r)$	average distance between test and reference signal
$d_i(t,s)$	distance between test and reference signals for frame i
$E[\bullet]$	expectation operator
E_i	energy in band i
f	normalized frequency (by the sampling frequency)
F	absolute frequency
F_0	pitch frequency
f_0	normalized pitch
F_m	maximum voiced frequency
F_s	sampling frequency
$\mathbf{f}_{UV}$	set of transformation functions for unvoiced frames
$\mathbf{f}_{VO}$	set of transformation functions for voiced frames
$G(n)$	LPC filter gain
GCI(i)	location of GCI i
$g_k(\mathbf{x}_n)$	Gaussian k
int($\bullet$)	nearest integer operator
N_0	number of samples in pitch period
PI	performance index
$p(C_i \mid \mathbf{x}_s)$	probability of $\mathbf{x}_s$ belonging to class i
$P(k)$	the power spectrum of the frame
$p(\mathbf{x} \mid \boldsymbol{\theta})$	pdf to be estimated using the parameters set $\boldsymbol{\theta}$
$p(c_k \mid \mathbf{x})$	probability that $\mathbf{x}$ is produced by class c_k
$p_i(c_k \mid \mathbf{x})$	probability that $\mathbf{x}$ is produced by class c_k in iteration i
$p(X \mid \lambda_A)$	probability of stimulus X belonging to speaker model λ_A
$P_{s,i}$	pitch of source speaker for frame i
$P_{t,i}$	pitch of target speaker for frame i
$\mathbf{s}$	column vector of the input speech
$\hat{s}(n)$	synthesized speech signal
$\hat{s}_h(n)$	synthesized harmonic part

$\hat{s}_n(n)$	synthesized noise part
$S(k)$	DFT of input speech
$S_h(k)$	DFT of harmonic part
$\hat{\mathbf{s}}_{\mathrm{h}}$	synthesized harmonic part of speech
$s(n)$	input speech signal
$s_h(n)$	harmonic part
$s_n(n)$	noise part
$S(f)$	log spectral magnitude function
$\mathbf{W}$	diagonal matrix having diagonal elements from Hamming
$w(n)$	Hamming window
w_k	mixture weights
μ	mean
$\boldsymbol{\mu}$	mean vector
σ	standard deviation
$\phi_l(n)$	phase of harmonic l at sample n
$\phi_{l,i}(n)$	interpolated phase for frame i and harmonic l
$\hat{\phi}_{l,i}$	actual unwrapped phase for frame i and harmonic l
γ	GCI perturbation control factor
$\boldsymbol{\theta}$	collection of parameters for m-Gaussian function
$\boldsymbol{\theta}_x$	GMM models of source speaker
$\boldsymbol{\theta}_y$	GMM models of target speaker
$\boldsymbol{\Sigma}$	covariance matrix
$\boldsymbol{\Sigma}_k$	covariance matrix of Gaussian k
$\boldsymbol{\Sigma}_{k,i}$	covariance matrix of Gaussian k iteration i
$\boldsymbol{\Sigma}_{xk}$	covariance matrix of source set for kth Gaussian

LIST OF ABBREVIATIONS

ACR	absolute category rating
ANN	artificial neural networks
AS	analysis-synthesis
CCA	canonical correlation analysis
CVC	consonant-vowel-consonant
CVQ	conditional vector quantization
DCR	degradation category rating
DCT	discrete cosine transform
DFT	discrete Fourier transform
DFW	dynamic frequency warping
DMOS	degradation mean opinion score
DTW	dynamic time warping
DWT	discrete wavelet transform
EGG	electroglottogram
EM	expectation minimization
EP	estimated phase
FIR	finite impulse response
GCI	glottal closure instant
GMM	Gaussian mixture model
HMM	hidden Markov model
HNM	harmonic plus noise model
IFC	inverse filter control
ISTFT	inverse short-time Fourier transform
LAR	log area ratios
LF	Liljencrants-Fant
LLR	Log-likelihood ratio
LMR	linear multivariate regression
LP	linear predictive
LPC	linear predictive coding
LSD	log spectral distance
LSE	least squares error
LSF	line spectral frequency
MAP	maximum a posteriori
MFCC	mel frequency cepstrum coefficient
ML	maximum likelihood
MLM	multivariate linear modeling
MMSE	minimum mean square error
MOS	mean opinion score
MQ	multivariate quadratic
MQM	multivariate quadratic modeling
Ms-LT	mixtures of linear transform
MVC	multistep-speaker voice conversion
PB	phonetic balance
PDF	probability density function

PESQ	perceptual evaluation of speech quality
PI	performance index
PPCAs	probabilistic principal component analyzers
PSOLA	pitch synchronous overlap add
PWD	perceptual weighting based distance
RBF	radial basis functions
RBFN	radial basis function networks
SD	standard deviation
SOLA	synchronized overlap and add
SOLAFS	synchronized overlap/add fixed synthesis
SP	source phase
STASC	speaker transformation algorithm using segmental codebooks
STFT	short-time Fourier transformation
STRAIGHT	speech transformation and representation using adaptive interpolation of weighted spectrum
TD-PSOLA	time-domain pitch synchronous overlap and add
TTS	text-to-speech
TVF	time variant filtering
UBM	universal background model
UL	univariate linear
ULM	univariate linear modeling
VFS	vector field smoothening
VQ	vector quantization

Chapter 1

INTRODUCTION

1.1 Problem overview

Voice conversion modifies the speech signal of one speaker (source) to make it perceptually similar to that of another speaker (target) [1]-[5]. It has a wide range of applications, such as voice verification systems [6], [7], speech enhancement [8]-[10], low bit-rate speech coding [11], cross-language speaker conversion [12], motion picture dubbing [13], cellular applications [8], interpreted telephony [8], text-to-speech systems [14], [15], speech compression [16], [17], and foreign language learning [18]. It is generally carried out using a speech analysis-synthesis system. It involves two phases: (i) estimation of source-to-target mapping or transformation function between the signal parameters derived from a set of phrases spoken by source and target speakers, and (ii) application of the estimated transformation function for conversion of the source speech [1], [14], [19]-[21]. For conveying speaker identity, the spectral parameters are considered to be relatively more important than those related to rhythm and intonation [19], [20], [22]-[27]. For voice conversion, several parameters have been used for representing spectral information: formant frequencies [28], [29], cepstrum [30]-[34], mel frequency cepstrum coefficients (MFCCs) [14], [20], and line spectral frequencies (LSFs) [24], [35]-[38].

The analysis-synthesis for voice conversion is generally carried out by segmenting speech signal into frames. The set of source or target parameters for each frame of the speech signal is known as a feature vector. The techniques for estimating the transformation function from the source feature vectors to the corresponding target feature vectors are generally based on vector quantization (VQ) [39]-[42], artificial neural networks (ANN) [29], [43], [44], mixtures of linear transform (Ms-LT) [45]-[47], hidden Markov model (HMM) [36], [48]-[50], Gaussian mixture model (GMM) [51]-[56], frequency warping [57]-[60], speaker interpolation [61]-[63], vector field smoothing (VFS) [43], [64]-[67], and time-variant filtering (TVF) [43], [64], [68]-[70]. Vector quantization suffers from the discrete nature of the acoustic space, which hampers the dynamic character of the speech signal. The statistical and ANN based techniques capture the natural transformation function independent of the

acoustic unit, but they need a large set of training data and computations. In frequency warping and interpolation, the transformation function can be estimated using lesser data, but a different transformation function is needed for each acoustic class. Vector field smoothing technique is effective only when it is clubbed with other statistical techniques and hence increases the complexity. In voice conversion by TVF, errors in estimating the filter parameters result in audible distortions.

1.2 Research objective

The research objective is to investigate the modification of spectral characteristics for voice conversion by modeling the relationship between the acoustic spaces of the source and the target using a single transformation function. Our hypothesis is that such a function applicable to all acoustic classes may be derived using multivariate polynomial modeling. Each parameter for generating the target speech is modeled as a multivariate polynomial function of the parameters of the source speech. The set of these functions is obtained from the time aligned source and target feature vectors. Harmonic-plus-noise model (HNM) has been used as the analysis-synthesis platform, as it provides high quality speech output with a reasonable number of parameters, and easily permits time and pitch scaling [71], [72]. As the HNM parameters (harmonic magnitudes and LPC coefficients) are not suitable for multivariate polynomial modeling, the harmonic magnitudes in the harmonic band are converted to MFCCs and the LPC coefficients in the noise band to LSFs for estimating the transformation function.

Voice conversion of the source speech signal is carried out by applying the estimated mappings for modification of spectral characteristics along with pitch and time scaling. Pitch scaling is used to match the range of the pitch in the source speech to that in the target speech. The pitch contour is modified without disturbing the duration of the speech signal. Time scaling is used to approximately match the duration of the source speech to that of the target. The duration is modified, with a scaling factor equal to the ratio of the total duration of voiced segments of the source to that of the target, keeping the pitch contour intact.

The technique is applied for same-gender and cross-gender voice conversion using parallel speech data for training. Evaluation is carried out using objective measures and listening tests.

1.3 Thesis outline

The second chapter gives a review of the techniques for voice conversion. The next chapter describes the HNM based analysis-synthesis platform for voice conversion and its implementation along with time and pitch scaling. The proposed voice conversion system for

modification of spectral characteristics and its implementation are presented in Chapter 4. The evaluation methods and results are presented in the following chapter. The last chapter provides a summary of the investigations and conclusions along with suggestions for further work.

[blank]

Chapter 2

VOICE CONVERSION

2.1 Introduction

For estimation of the source-to-target mapping or transformation function, the speech signal is analyzed for extracting non-linguistic or speaker-specific information [73]. Speaking style and vocal quality are the two important components of the non-linguistic information. Speaking style depends upon pitch contour, duration of words, timing, rhythm, pause, power levels, etc. Vocal quality depends on physiological properties of the glottal source and vocal tract [74]-[77]. The vocal quality may be described using the short-term parameters such as shape of spectral envelope and spectral tilt, formant frequencies and bandwidths, formant transitions, long-term average speech spectrum, and jitter [75], [76], [78]-[87]. The speaker has relatively less control over these parameters, and they are considered as important indicators of identity of the speaker [18], [88]-[91].

The transformation function may be estimated using a text-dependent (parallel data) system [1], [4], [5] or text-independent (non-parallel data) system [2], [3], [45], [92]. In text-dependent scheme, the utterances from the source and the target speakers correspond to the same written text. The text-independent scheme does not pose any such constraint. The parallel data systems are relatively more efficient in estimating the transformation function as frame-by-frame alignment of the source and target parameters preserves the phonetic context [37], [93], [94],. The alignment of the parameters is usually carried out by dynamic time warping (DTW) [20], [95], [96].

The commonly used spectral parameters for estimating the transformation function are formant frequencies [28], [29], cepstrum [30]-[35], mel frequency cepstrum coefficients (MFCCs) [14], [20], and line spectral frequencies (LSFs) [24], [35]-[38]. MFCCs and LSFs are considered more suitable for voice conversion. MFCCs use mel scale based compression of the spectral parameters which reduces the perceptual effect of the errors introduced by the discretization of the spectrum [97]. Despite spectral smoothening, the coefficients do not loose high frequency information and they are uncorrelated to each other [98]. The corresponding coefficients in source and target MFCCs have been reported to be correlated,

and this property is very useful for using them in stochastic modeling [20], [99]. They have also been reported to be robust with respect to noisy environment [100]. The main features of LSFs are that they have a definite relationship with formant frequencies and bandwidths and they can be reliably estimated. They have a limited dynamic range and show good linear interpolation properties. Further, effect of errors in any LSF during voice conversion is localized [94], [101], [102].

The estimated transformation function from the spectral parameters is subsequently applied on the source parameters to obtain the target parameters. For voice conversion, the different modifications required may be categorized as spectral envelope modification, excitation or residual modification, and prosodic modification. The techniques for estimating the transformation function for spectral envelope modification are generally based on vector quantization (VQ) [39]-[42], artificial neural networks (ANN) [29], [43], [44], Gaussian mixture model (GMM) [51]-[56], mixtures of linear transform (Ms-LT) [45]-[47], hidden Markov model (HMM) [36], [48]-[50], frequency warping [57]-[60], speaker interpolation [61]-[63], vector field smoothing (VFS) [43], [64]-[67], and time variant filtering (TVF) [43], [64], [68]-[70]. These techniques are described in the following sections. The techniques for the residual and the prosodic modifications are presented in Section 2.8 and Section 2.9, respectively. The last section gives a summary of this chapter.

2.2 Vector quantization

In vector quantization (VQ) based voice conversion, the acoustic spaces of source and target speakers are partitioned into finite classes. A mapping is established between the corresponding mean feature vectors of these classes using histograms. The mapping is stored in the form of a code book. Shikano *et al*. [39] and Abe *et al*. [40] used a VQ codebook of 256 spectral LPC (order =12) feature vectors. Two other scalar codebooks were used for the pitch and power transformation. The codebooks were designed using the K-means algorithm [103]. Ten phonetically balanced words were used for training. The transformation function was obtained by piecewise linear mapping between the codebooks of the source and the target speakers. Transformations resulted in a significant reduction in the log spectral distances. Subjective evaluation using AB test showed about 60% correct responses for the transformed voice to be similar to that of the corresponding target.

Abe [41] used a segment based approach to capture dynamic characteristics of the speaker individuality. The speech signal of the source speaker was segmented into phonetic units using a speech recognizer. The source phonetic units were replaced with corresponding units of the target speaker using a table-lookup approach. The quality of the output speech was low as the converted speech was synthesized using a limited number of phonetic units [28], [73]. Knagenhjelm and Kleijn [104] observed that spectral discontinuities because of

phonetic units replacement degrade the quality. Stylianou *et al.* [20] reported that VQ based methods [34], [40], [116], [57], introduce discontinuities during segment transitions leading to degradation of the quality of the transformed speech. As a solution, soft-clustering approaches have been suggested [20], [31], [105].

2.3 Artificial neural networks

Artificial neural networks (ANN) have the ability to estimate the transformation function involving even a high degree of complexity. Narendranath *et al*. [29] used an artificial neural network for transformation using formant based analysis-synthesis. A feedforward neural network with one input layer, two hidden layers, and an output layer was trained using back propagation algorithm for transforming formants of the source speaker to that of the target speaker. Evaluation using vowels showed a need for improvement in the quality. This may be due to the problems of slow convergence, low data resolution, and local minima traps in back propagation modules [106]. ANN based on radial basis functions (RBF) have the ability of estimating transformation function with higher learning speed and higher clustering abilities [106]. But they are sensitive to the initial value and sometimes fail to converge due to local minima. To address these shortcomings, Zuo and Liu [44] used genetic algorithm to train the hidden layer of RBF network for enhancing the ability for global optimization. The output weights of RBF network were estimated using gradient descent method. The perceptual distance measures did not show improvement for the vowel transformation. Chen and Zhang [107] reported that the use of joint parameters of LSFs and pitch for training the ANN improved the results.

2.4 Gaussian mixture model

In Gaussian mixture model (GMM) based systems [51]-[56], [101], [108], [109], the probability distributions of acoustic parameters of the source and target speakers are modeled by a finite number of Gaussian functions using maximum likelihood (ML) or expectation maximization (EM) criteria. This technique minimizes the effect of textual differences between the training and test utterances, and it does not need explicit speech segmentation. In voice conversion systems using GMM, it is assumed that the speech signal consists of a finite number of acoustic classes (such as vowels, nasals, fricatives, etc.) and each class is characterized by an average spectral feature vector along with some variability because of pronunciation and co-articulation effects. The distribution of speech parameters is represented as a weighted sum of a finite number of multivariate Gaussian functions, with each function specified by its mean feature vector and the covariance matrix accounting for the variability

around the mean feature vector. The transformation function consists of a linear mapping between the parameters of the Gaussian mixture models of the source and target speakers.

Stylianou *et al.* [20] employed GMM for obtaining transformation function in harmonic plus noise model (HNM) based analysis-synthesis framework and demonstrated the superiority of GMM over codebook methods. The GMM parameters were estimated by minimum mean square error (MMSE) criterion [14], [24]. The maximum voiced frequency, which separates the harmonic and noise bands in the speech frame spectrum, was fixed at 4 kHz. The harmonic magnitudes were converted to MFCCs (20 coefficients). The noise part was transformed by using two 6th order LPC correction filters, one for the voiced frame and the other one for the unvoiced frame. Source phase was used for synthesis. A set of manually segmented and labeled 1500 diphones of French language were used for training. The prosody was modified by average pitch and time-scaling. Cepstral distance was used for objective evaluation. The investigations showed that the cepstral distance was reduced by more than 4 dB between the source and the target frames. The conversion between two male speakers was evaluated by conducting XAB, preference, and opinion tests, using three utterances and 21 listeners. It showed that 97% stimuli were correctly identified by the listeners. The preference test score for GMM (64 components) in comparison to VQ was 71.8%. Opinion test confirmed that the transformed-target distance was lower than the source-target distance. The overall quality of the converted signals was reported as satisfactory although some of the listeners reported a muffling effect when the number of GMM components was smaller than 64. Erro *et al*. [110] investigated a similar transformation system with a variant of HNM using constant frame rate instead of a pitch-synchronous scheme. GMM based transformation function was estimated by converting the HNM magnitudes to LSFs. Evaluation using preference test (18 listeners, 17 Spanish sentences) showed the system to be better than TD-PSOLA (time-domain pitch synchronous overlap and add).

Toda *et al*. [111] addressed the problem of muffling using global variance based ML criterion and dynamic parameters generation algorithm [112]. MFCCs (24 coefficients) were used to estimate the GMM (128 components) based transformation function using fifty manually segmented and labeled sentences for one male and one female speaker sampled at 16 kHz. Two transformation functions (one for male-to-female and other for female-to-male) were estimated. The synthesis was carried out by using STRAIGHT (speech transformation and representation using adaptive interpolation of weighted spectrum). Evaluation using MOS test (25 sentences, 5 listeners) showed that the score became 3.25 from 2.25 using global variance as compared to ML with almost no change in the opinion score.

Najjary *et al*. [113] investigated the effect of joint distribution of pitch and MFCCs (20 coefficients) using GMM (64 components) on the quality of the transformed speech. Pitch

was modified by using mean values and standard deviations of pitch frequencies of the source and the target speakers. The training was performed with 30 short-paired utterances (16 kHz) from one male and one female speaker, with manually corrected segmentation. Preference test (10 utterances, 10 listeners) showed that the speech modified by joint distribution of pitch and MFCCs was preferred in 97% cases.

Percybrooks and Moore [18] investigated the effect of excitation using a pitch-synchronous LPC based voice conversion. The source residual was used for unvoiced and average target excitation for voiced frames; with GMM based mapping trained using 35 sentences from VOICES [24] in LSF domain. Evaluation using spectral distance measure, MOS, and XAB (15 sentences, 8 listeners) showed that inclusion of the residual estimation made the converted speech closer to the target voice, but it also introduced a higher distortion due to phase discontinuities of residual representation.

Gao and Yang [90] used a GMM (64 components) based transformation trained by aligned LSF vectors obtained from 10 short sentences (16 kHz) in Mandarin Chinese (2 male and 2 female speakers). The pitch of the source residual was modified by using mean and standard deviations of pitch frequency of the source and target speakers. For synthesis, LPC based platform was used. To enhance the quality of the speech, a perceptual filter was also applied to the transformed speech. Subjective test using XAB for a male-to-male transformation resulted in scores of 78.4% and 80.2% for VQ and GMM based systems, respectively.

Dutoit *et al*. [91] compared three voice conversion techniques. In the first one, transformed speech was obtained by source residual and actual target MFCCs (20 coefficients), aligned by Viterbi algorithm. In the second case, source residual and the nearest target feature vector to GMM (256 components) transformed vector using Viterbi alignment were used for obtaining the transformed speech. In the third case, target excitation aligned by DTW and GMM (256 components) transformed vectors were used. The target speech was generated using LPC (order = 20) in all the three cases. Evaluation using opinion test (10 listeners) showed that the third method provided best results with respect to similarity but worst quality.

Zhao and Gao [114] investigated the effect of speaking rate on the transformation. The speech from the source speaker was first modified by synchronized overlap/add fixed synthesis (SOLAFS) to match its rate to that of the target speaker. This was followed by estimation of GMM (20 components) based transformation function in LSF (order = 18) domain. Pitch modification was carried out by frequency domain compression or expansion of the source residual by pitch scaling on real and imaginary parts using spline interpolation followed by inverse short-time Fourier transform (ISTFT). For enhancing the quality, a perceptual filter was used [115], [116]. The system was trained using Mandarin Chinese (3

male and 3 female speakers, 16 kHz). Itakura spectral distance measure showed that average reduction ratio increased from 43.8% to 62.2% and XAB score increased by 12% for a male-to-male transformation.

Mouchtaris *et al*. [42] attributed the reason for muffled quality of the GMM transformed speech to averaging of the one-to-many relationships between the source and the target speakers. It was suggested that the quality can be improved by using conditional vector quantization (CVQ). This method picks up the actual target vector instead of averaging one to many instances. Joint probability density of both the source and target speakers has also been used for improving the quality of the transformed speech [24], [94], [117]. Although this method increases the complexity due to EM, it obviates the need to perform MMSE. Modeling the joint probability density allows the system to capture all possible correlations between the source and the target speaker spectra. MMSE estimation assumes the feature vectors to be independent. As the variance of each component of the transformed feature vectors is not considered, it loses second order statistical information [14], [37], [118]-[121]. As LPC residual does not contain any significant second order relations corresponding to the shape of the vocal tract [121], Choi and King [118] reported that canonical correlation analysis (CCA) estimation provided a better performance than MMSE [118], [122] by preserving second order information. Jian and Yang [45] also investigated the use of CCA for estimating the transformation function. Source residual was pitch modified using means and standard deviations of the source and target speakers before LPC synthesis. The system was trained by using 18 Mandarin Chinese sentences (16 kHz) from 2 male and 2 female speakers. XAB test (8 listeners) showed about 2% increase in the scores, but almost no difference was observed using Itakura spectral distance measure.

Lee [105] derived the GMM based transformation function using cross correlation probabilities of DTW [123] aligned LPC cepstral parameters (order = 30). Synchronized overlap and add (SOLA) [124] algorithm was used for modifying the speaking rate (average vowels/s) of the source residual. The pitch was modified by average scaling factor. Modification of excitation signal was carried out by linearly interpolating the real and imaginary parts of the short-time Fourier transformation (STFT). The training was performed by 20 Korean sentences (16 kHz) from three male and one female speakers. Evaluation using spectral distance, LLR, XAB (15 sentences, 18 listeners), and preference test (10 sentences, 15 listeners) showed some improvement, but some of the converted sentences were found to have seriously degraded intelligibility, naturalness, and identity.

Jian and Yang [45] used a GMM based system involving mixtures of linear transform (Ms-LT) to model the source and target feature vectors, and reported that this technique avoided the need of parallel training data. The coefficients of the linear transformation in LSF (order = 16) domain were obtained by expectation-maximization. The over smoothening of

the formants was compensated by chirp Z-transform [125]. The synthesis was carried out using LPC platform after modifying the pitch of the source residual by mean and standard deviation based method. A total of 200 syllable-balanced (16 kHz) Mandarin Chinese utterances from two male and one female speaker were taken for training. Evaluation used Itakura spectral distance measure and XAB test (8 listeners) for fifty utterances. The authors reported that their system provided results comparable to that of conventional systems using parallel data [24].

Masuda and Shozakai [126] introduced the concept of multistep-speaker voice conversion (MVC) for reducing the complexity of estimating a separate transformation function for each pair of speakers by introducing an intermediate speaker and computing the transformation function from each speaker to the intermediate speaker. GMM (64 components) based transformation function was obtained from cepstral coefficients (41 coefficients) extracted by the STRAIGHT analysis method [127], [128] from fifty phonetically balanced Japanese sentences (16 kHz) from three male and three female speakers. Two speakers were taken as source and target and four as intermediate speakers. Experimental results based on cepstral distance and DMOS (degradation mean opinion score) test (3 sentences, 5 listeners) showed that the speech quality of the converted speech was comparable to that of conventional systems.

In general, GMM can model even complicated dependencies between variables if the size of the training data and the number of mixture components are not limited. Estimation of the transformation function using limited data leads to improper mapping due to over fitting and over smoothening [99], [129]. Further, GMM based methods assume frame-to-frame time independence and this introduces some degradation in the speech quality [19].

2.5 Hidden Markov model

Voice conversion can be carried out by partitioning the acoustic spaces of the source and the target in equivalent classes instead of independent feature vectors. This approach preserves the natural relationship between the consecutive frames and modeling of the frames of the individual phonemes or diphones after segmentation provides better results [130], [131]. Segmentation of the input speech in phonemes or diphones can be achieved by using hidden Markov model (HMM) [20], [35], [36], [132], [133] and Viterbi algorithm [95]. The state transition property in HMM based methods presents a good approximation of the spectral envelope evolution along the time axis. The HMM is trained with the source and the target speech data simultaneously. It models the probability distribution of the feature vector sequence according to its actual state sequence and the transition probabilities between the states.

Arslan [35] reported the performance of sentence HMM based STASC (speaker transformation algorithm using segmental codebooks) better as compared to phonetic STASC for source-filter based voice conversion, especially for nasalized sounds. Source and target codebooks were generated using LSFs (20 order, 16 kHz). The ratio of target and source spectra was used as the vocal tract transfer function. These spectra for input frames were constructed from the weighted centroids. The weights were determined by the distances among input frame feature vectors and the centroids of the classes. Same approach was used for estimating the target excitation spectrum. This method was able to transform not only general excitation characteristics, but zeros as well. Prosody was modified by codebook approach and the evaluation (using 2 utterances from 2 male and 1 female speakers) was carried out using cepstral distance, speaker identification, XAB, and intelligibility tests. It was reported that the performance was context dependent and degradation increased when the source and the target speakers were very different.

Arslan and Talkin [49] employed a sentence HMM using 18 order LSF vectors in STASC [35] approach. A left-to-right HMM with no skip state was trained for the source utterances and the target utterances were force-aligned with the automatically labeled source utterances. The number of states for each utterance was directly proportional to the duration of the utterance. A new state was added to the HMM every 4 ms. With this model, neither the text nor the language of the utterance needed to be known. To compensate the energy and speaking rate differences between two speakers, a codebook based duration and energy scaling algorithm was proposed. The target excitation was obtained from that of the source using mean and standard deviation. Subjective listening tests using three subjects showed that intelligibility was maintained at the same level as natural speech after the voice conversion using fifteen short nonsense sentences. This method has also been explored using discrete wavelet transform (DWT) [134].

Salor *et al*. [50] used HMM for estimating the finite impulse response (FIR) filter based voice conversion in LPC framework. The pitch contour of the target was modified using a time-domain pitch-synchronous overlap and add (TD-PSOLA) method. Training was performed by 35 phonetically balanced sentences (16 kHz) from two male and three female speakers, from METU Turkish database [135]. Performance using log-likelihood ratio (LLR) and five sentences showed that LLR increased from -6.61 to 0.92 after transformation.

Qin *et al*. [136] used STASC for STRAIGHT (speech transformation and representation using adaptive interpolation of weighted spectrum) [137] based voice conversion. The glottal formant estimated by first Gaussian component of GMM model from both source and target spectra were first removed to make the spectrum excitation independent. The modified spectrum was modeled by an all-pole model with 20 LSF coefficients. Source LSFs were converted to target LSF by codebook mapping. Based on the linear relationship between pitch

frequency and glottal formant, target spectrum was modified to add the glottal formant. The training was performed using 10 sentences. The evaluation using MOS test (10 sentences, 5 subjects) showed improvement as compared to STASC.

Verma and Kumar [138] employed a voice fonts (spectral envelope, fundamental frequency, and speaking rate) based voice conversion using HNM platform. The source and target sentences aligned by HMM and Viterbi techniques were divided into 61 classes. Each class was modeled using GMM (128 components) and the transformation function was obtained by assuming each target class as a function of five most likely classes in the source acoustic space. The harmonic magnitudes were converted to MFCCs (16 coefficients). The noise band was modeled by LSFs (order = 16). The time scaling was carried out by average durations of various acoustic categories stored in the voice fonts. Pitch was modified by using mean and standard deviations of each class. The system was trained by using fifteen Hindi sentences (16 kHz) from six male and four female speakers. Evaluation using spectral distance, absolute category rating (ACR), and degradation category rating (DCR) with 45 transformed sentences and 10 listeners showed a marginal improvement over conventional GMM based technique.

Turk and Arslan [37] addressed the problems of quality degradation due to differences in speaker characteristics, recording conditions, and signal processing algorithms using confidence measures, pre-emphasis, and spectral equalization. The aligned frames of the source and target speakers having differences of parameters (spectral distance, pitch, energy distance, and duration difference) greater than 1.5 times the standard deviation from the mean were discarded before training. For spectral equalization, vocal tract transformation function was multiplied by smoothened (along frequency) ratio of target and source long-term average power spectrum. Training used sixteen Turkish sentences (44.1 kHz) from two male and two female speakers with STASC approach. Evaluation using 10 listeners showed that the proposed algorithm was preferred over the baseline algorithm by 76.4%, improved the similarity to the target voice by 23.0%, and enhanced the MOS test score by 46.8%, respectively.

Modeling source and target speech in a joint HMM may introduce confusion in the mixture densities and the transition probabilities. In a standard HMM, the state duration probability decreases exponentially with time, which is inappropriate for some of the speech segments [139]. To address this problem, a DeBi-HMM for the modeling of the source and the target speech has been reported [140]. Gamma distribution was embedded as the duration model for each state and quality of the synthesized speech was reported to be satisfactory.

Ye and Young [36] investigated the effect of phases on the quality of transformed speech and attributed the harsh quality of the speech due to source phase. The target phases were estimated by classifying the target LSFs into 64 GMM based classes. Posterior

probabilities of a frame to be a member of these classes was taken as a weighted vector set, with the weights derived using the method of least squares error. The phases of the resulting predicted waveform were used in the synthesis of the target speech. This phase prediction method was compared with two phase coding methods involving the minimum phase and the phase codebook approaches [125]. The residual (difference of log actual magnitude spectra and LPC envelope) was also predicted from the same approach. The unvoiced frame sequences labeled by HMM were selected from the target database. The extracted target frames were modified in amplitudes to match the source frames. The synthesis was carried out by pitch-synchronous sinusoidal model. Evaluation using Itakura spectral distance and XAB (23 listeners, 32 sentences) showed that this approach outperformed the other two methods with respect to identity, but spectral distortion and many other artifacts were present in the output even after using perceptual filter [115], [141].

Jian and Yang [131] used a Viterbi algorithm [95] approach for aligning the source and target feature frames. GMM was trained from the aligned speech in LSF domain using 18 Mandarin sentences (2 males and 2 females, 16 kHz). XAB score was reported as increased by 2% as compared to simple GMM based system, although Itakura spectral distortion ratio remained almost comparable for both.

Bandoin and Stylianou [142] investigated four techniques for voice conversion namely, VQ [40], GMM [143], ANN [29], and linear multivariate regression (LMR) [57] using French CNET database (16 kHz). Evaluation using two sets of phonetically balanced sentences with cepstral distance measure showed that the GMM based technique resulted in the largest reduction in the transformed-target distance. The performance of the ANN and LMR based voice conversion was almost same but better than that of VQ based voice conversion.

2.6 Frequency warping

In dynamic frequency warping (DFW), the mapping between the log-magnitude spectra after removing the spectral tilt of the source and the target speaker is estimated. The number of warping functions is equal to the number of source-target pairs of spectral feature vectors within the class. The total warping functions may be averaged for getting a single transformation function for each acoustic class. Valbret *et al*. [57] modeled the log-magnitude spectrum using a discrete set of frequencies [144], [145] and the spectral distortion between pairs of DFT indices was computed. DTW was used to find a path between the indices of the source and target spectra for minimizing the spectral distortion. The average warping function for each class was modeled as a third degree polynomial. For comparison, linear multivariate regression (LMR) was also used to estimate the transformation function. The synthesis was performed using source residual, with the prosody modified by PSOLA and LPC parameters extracted from the transformed spectrum. Training was carried out by CVC syllables (16

kHz) with 10 vowels and consonants from the set /b, d, g, p, t, k/ from four male speakers. Eight instances of each syllable were extracted (6 for training, 2 for testing). XAB test with three listeners and three syllables showed that LMR performed better than DFW with some audible distortions. On the contrary, DFW speech sounds were reported as smoother but the transformed speech was perceived as being in between the source and the target. The evaluation using sentences was reported as unsatisfactory. Similar conclusion about the quality were drawn for multi-segment based frequency warping functions [58], generally used in the context of vocal tract normalization [15], [146] for cross-language voice conversion [58], [147].

Toda *et al*. [59] investigated GMM based (64 components) algorithm with dynamic frequency warping by frequency-dependent weighting to avoid over smoothening. Frequency warping function was estimated as the path which minimized the normalized spectral distance between the STRAIGHT log spectrum of the source speaker and the target speaker in 2-dimensional frequency index plane. To convert the spectral power, the frequency-weighted residual spectrum (difference between the GMM based converted log spectrum and the dynamic-frequency-warped log spectrum) was added to the converted one. Prosody was modified by average of log-scaled fundamental frequencies. Training was performed using 58 sentences. XAB test (8 listeners) and cepstral distance measure showed improvement in the scores. In [148], it was reported that the similarity score for the transformed speech could be improved by 20% with respect to DFW if the source magnitude spectrum below 100 Hz is left unaltered for maintaining the continuity between consecutive frames.

The estimation of DFW based transformation function may be simplified by using the mapping between formants instead of spectral envelope. Slifka and Anderson [60] used mean and standard deviation of the angle and the radius of the pole locations corresponding to the formant resonances for two vowel classes of the source and the target speakers to match the target speaker mean and standard deviations. The listening tests showed that apparent speaker identity was altered, but the target identity was not achieved. It could be attributed to the fact that the speaker individuality is also affected by the voice source [73]. Ueda *et al*. [140] transformed five Japanese vowels using source and target formants. The formant tracking was carried out by the inverse filter control (IFC) as reported in [149]. The source and target spectra were normalized by warping each formant position such that they were equidistant. The normalized spectra at any point in F1-F2 space was obtained by interpolation or extrapolation of the nearest source spectra. The inverse mapped spectrum obtained from the target formants was used to obtain the minimum phase impulse response. Converted speech signal was synthesized as the convolution of the pulsed excitation and this impulse response. Opinion test showed that cross-gender conversion was not satisfactory.

Arriola *et al*. [150] transformed speaker dependent parameters (gain contour, pitch contour, glottal source parameters (in Liljencrants-Fant (LF) model), and vocal tract parameters (formants and bandwidths)) by linear regression based transformation function estimated for each parameter using DTW. Synthesis was carried out by pitch-synchronous formant synthesizer for two male and one female voice with manual adjustment for pitch and formants. The analysis of converted speech showed conversion but the quality was not satisfactory. Mizuno and Abe [28] proposed piecewise linear conversion rules controlling formant frequencies, formant bandwidths, and spectral intensity to produce speech with the desired formant structure using a codebook of 256 entries. XAB and preference tests were used for evaluating the quality of the converted speech. Three words of one male speaker were converted to the speech of other three male speakers. The output of this method was compared with the output of VQ method, which provided slightly better results. In general, the quality of the formant based voice conversion is low because of errors in the automatic estimation of positions, bandwidths, and amplitudes of the formants [28], [52], [57], [151].

2.7 Speaker interpolation

In speaker interpolation, different styles in the synthesized speech are generated by mixing different voices of the same or different speakers using interpolated and weighted spectra [61]-[63]. For example, in [61], [152], spectral parameters (12-order log area ratios and 30-order LPC-cepstrum) were used to represent each target frame as a weighted sum of corresponding frame spectra of the stored data from 4 speakers. The interpolated feature vector was given by

$$\hat{\mathbf{y}}_i = \sum_{n=1}^{N} w_n \mathbf{x}_{n,i} \tag{2.1}$$

where $\sum_{n=1}^{N} w_n = 1$. Here $\mathbf{x}_{n,i}$ represents the feature vector for frame i in a time-aligned utterance of speaker n, N is the number of pre-stored speakers, w_n is weighting coefficient (interpolation ratio), and $\hat{\mathbf{y}}_i$ represents the interpolated spectral feature vector for frame i. The values of the interpolation ratios were determined for minimizing the error between the interpolated and target feature vectors. The weighting coefficients estimated from the training were used to synthesize the target speech from the stored speech units of the source speakers. Evaluation using Japanese sentences showed that the cepstral distance between the target speaker and that of transformed speaker was reduced by about 25% than the original source / target distance. It was reported that the formants in the transformed speech were broadened due to over smoothing and hence the conversion system needed to be refined.

In vector field smoothening (VFS) [43], [64], the correspondence between feature vectors from different speakers is assumed to be a smooth vector field. The average difference from the desired target to the nearest target is found for each acoustic class and the conversion is achieved by adding this difference to the feature vectors of the nearest target. This scheme needs relatively less data [64], but the important dynamic parameters of speech might be lost due to smoothening. Voice conversion based on time variant filtering (TVF) uses a few short adaptation units (phonemes or short words) [43] and the voice conversion is performed using a time-variant digital filter. The filter coefficients of the time-variant filter are selected by the feature map dependent on the short-time spectrum. Unlike other voice conversion systems, this system transforms the speech signal instead of the parameters, but a reliable estimation of the filter coefficients is difficult.

2.8 Residual modification

The modeling error in speech analysis, the difference between the synthesized speech and the original speech is known as the residual. In LPC based framework, it corresponds to the excitation. In HNM based framework, it is the noise part. Although most of the information regarding the identity of the speaker is contained in the vocal tract parameters [25], some transformation of the residual as well is needed for an effective voice conversion [20], [23]-[27]. The different methods used for this purpose can be grouped as source residual, excitation parameterization, excitation conversion, excitation prediction, and excitation selection.

In the source residual method [94], the parameters of the residual of the source speaker are modified in accordance with that of the target speaker. However, the modification of the residue cannot be carried out independent of the spectral parameters of the vocal tract. The modeling errors present in the valleys of the spectra may affect the peaks and degrade the quality of the transformed speech [153], [154].

In excitation parameterization, the shape of the excitation is modeled using a small number of parameters. As the excitation can be assumed almost similar within each phonetic class [27], the total number of excitation patterns may be considered as finite. In Milenkovic's work [87], the derivative of the excitation obtained by inverse filtering the speech in each pitch period was modeled by a sixth degree polynomial. Childers [154] used clustering for developing a codebook of glottal wave derivatives using the coefficients of the polynomials estimated for each voiced frame using polynomial and LF model [155] for representing breathy, creaky, modal glottal waveforms. It was reported that one type of voice could be modified to another type by changing the model parameters. However, since coefficients of the polynomial have no physiological interpretation, feature vectors of the source speaker could not be converted to those of the target speaker directly. LF model for glottal derivative

could overcome this problem but high frequency information of glottal wave would have been lost. It has been observed that breathiness and roughness are the outcome of turbulent noise and aperiodicity in the glottal waveform. They affect the harmonic structure and spectral noise level of the speech spectrum. These effects cannot be modeled by using simple models like LF model [27]. Modeling excitation using a single pulse in voiced regions and white noise with random phases in unvoiced regions, may result in a speech with mechanical quality [21], [24]. Lee *et al*. [35] modeled the excitation by a long delay neural net predictor whose parameters were mapped based on the maximum occurrence in a 2D histogram of feature vector correspondences. In these attempts, the quality of the output could not be made satisfactorily natural.

In excitation conversion, the target excitation is estimated from the source residual [35], [156]. In [14], [127], [136], [137], the spectrum was decomposed into excitation-dependent and excitation-independent components. Each component was modified separately. Listening test showed some improvement in the quality of the transformed speech [136].

The modification of the residue may also help in precise estimation of transformation function. For this, the source speech is first modified to match the prosody of the source to that of the target using residue modification. This reduces the spectral differences in the source and target parameters. Lesser the spectral differences, better is the alignment using DTW [114]. Rao and Yegnanarayana [157] modified the prosody (pitch contour and speaking rate) in excitation domain using the instants of significant excitation, before alignment and reported some improvement. The instants of significant excitation correspond to the instants of glottal closure (epochs) in the case of voiced speech, and to some random excitations like onset of burst in the case of unvoiced speech [157]. Positive zero crossings of the filtered average (3-point median) group delay were taken as points of significant excitation [157].

Stylianou and Cappe [51] transformed the residual (noise part) using HNM based synthesis [158] by using two sixth order LPC corrective filters. The first filter is used to transform voiced segments and the other one is used for transforming unvoiced segments in the noise part. The coefficients of the filters are estimated from the difference of average spectral power densities of the source and target speakers.

The next method for residue modification is the excitation prediction. In this, the target excitation is estimated from the vocal tract parameters. It has been reported that LSFs have finite amount of correlation with the corresponding excitations and hence, excitation can be predicted from the vocal tract parameters [21], [24], [36], [159]. In each class, a weighted average of all target excitation magnitude spectra is stored. The weights are the posterior probabilities that a given feature vector belongs to a class. During the conversion phase, target excitation magnitude spectrum is calculated by a weighted sum of all target class excitations. To make the code-words pitch-independent, the original excitation feature vectors can be

upsampled to a common length using a nearest-neighbor interpolation scheme [24]. The excitation phases were smoothened after unwrapping in time using an 8-point FIR filter to reduce artifacts [24].

Sun *et al*. [27] extracted LSFs and glottal wave derivative samples (with normalized length and amplitude) in each pitch frame and divided them in classes using clustering algorithm. The excitations nearest to the centroids of the classes were also stored. At the time of voice conversion, the target excitation was found by the closest match of the vocal tract parameters with the centroids. Experimental results showed that this method significantly outperformed Rosenberg model and LF model [27].

Kuin and Macon [24] predicted target residual from the LPC cepstral parameters during voiced speech. In this method, LPC cepstral coefficients of voiced segments are clustered by a GMM with 32 classes. Each cepstral vector is associated with a residual complex spectrum (residual magnitude spectrum by subtracting the LPC log-magnitude envelope from the original log-magnitude spectrum, residual phase by the difference between the LPC system phase and the original phase spectrum). To make the code words pitch-independent, the original residual vectors are up sampled to a common length using a nearest-neighbor interpolation. For each class, the magnitude spectrum is calculated by a weighted mean of all magnitude vectors, corresponding to the normalized probability of belonging to that class; the phase spectrum is set to the centroid phase. The phases are unwrapped in time and smoothed by an 8-point FIR filter to reduce audible artifacts due to sudden changes in the residual phase. Finally, the residual spectrum is added to the LPC spectral envelope.

In [35], [160], separate filter was used for each phoneme class to transform the excitation. The transfer function for the filter was obtained using the average short-time magnitude spectra for each class

$$H(\omega) = \sum_{i=1}^{N} p(C_i \mid \mathbf{x}_s) \frac{U_{t,i}(\omega)}{U_{s,i}(\omega)} \tag{2.2}$$

where N is total number of classes, $p(C_i \mid \mathbf{x}_s)$ is the probability that the given source residual belongs to class i, $U_{s,i}(\omega)$ and $U_{t,i}(\omega)$ are source and target average magnitude spectra for class i. This filter was able to transform zeros of the residuals, which are not represented accurately by all-pole modeling and hence, showed improved quality for nasalized sounds. But jitter, aspiration, and noise bursts needed further transformation [160].

The residual selection technique stores all original excitations used during training into a table together with the corresponding feature vectors [21], [24], [36], [161]. The listening tests showed that there were some artifacts introduced due to insufficient correlation between feature vectors and excitations. These problems can be avoided by preprocessing the

excitations to reduce abrupt changes in the voiced regions and random behavior in unvoiced regions using random phases. The quality can be further improved by unit-selection based residual prediction [21], [162]. It has been reported that retaining the source residual provided relatively better speech quality, but residual prediction method provided better similarity of the transformed speech to the target speech [161].

2.9 Prosody modification

Prosody is a function of many parameters such as pitch, duration, and energy. It helps in interpreting the utterances by grouping words into larger information units and drawing attention to the specific words. Prosody varies with speaker's attitude, confidence, and mood in a conversation [19]. Speaking style of a speaker is highly correlated with prosody information [30]. Kawanami *et al*. [163] and Murray and Arnott [80] have reported that prosodic parameters provide useful cues in emotional speech. Hence, prosody modification is highly desirable for voice conversion [29], [30], [35], [40]. The different methods for estimating the target prosody can be grouped in three categories: rule based, VQ based, and statistical.

In rule based prosody modification, the current prosody is predicted from a set of rules. The rules are defined according to the context and the emotions to be conveyed. The rules are obtained manually after analyzing the different speech patterns of the speakers. Rule based prosody modeling approach has been used for prosody modification [164]-[166]. Silverman [167] has indicated that a domain-specific prosody model can significantly improve the comprehension of the synthesized speech. However, the design of these rules is labor-intensive and difficult. In [168], a text independent automatic accent classification system using phone based models and principal component analysis was reported for prosody adjustment.

VQ based methods store the prosody patterns of the source and the target in a codebook. For each source prosody pattern, the target prosody pattern is picked up from the codebook [41], [169]. Because the spectral parameters are not independent of prosody, a large data is required to capture all the possible prosody patterns [80], [163], [170]-[173]. Template tree [165] and decision tree [174] methods have been also proposed for prosody modeling.

In statistical methods for prosody modification, each prosody parameter (pitch frequency, duration, or energy) is obtained by linear transformation (2.3) [20], [44], [49], [90], [94], [124], [131], [138], [173]-[179]. For frame model *i*, the target parameter $P_{t,i}$ is obtained from the source parameters $P_{s,i}$ using

$$P_{t,i} = \mu_{t,i} + \frac{\sigma_t}{\sigma_s}\left(P_{s,i} - \mu_s\right) \tag{2.3}$$

where $\mu_s, \sigma_s, \mu_t, \sigma_t$ are the source mean, source standard deviation, target mean, target standard deviation, respectively. The statistics used may be phrase based or phonemic class based [35], [124], [173], [178], [179]. Instead of modeling the mapping on a linear scale, the linear transformation may be obtained on logarithmic scale [148],

$$\log P_{t,i} = a \log P_{s,i} + b \tag{2.4}$$

If the prosody is to be modified in excitation domain, the regions near the epochs should not be disturbed for minimizing the distortion in the transformed speech [157]. Precise estimation of the pitch is difficult, particularly for high pitched voices or noisy conditions [180], [181]. Improper pitch estimates deteriorate further speech processing and result in unpleasant speech quality. Hence, pitch contour smoothening is necessary to make the quality acceptable.

2.10 Summary

Several methods for voice conversion have been reviewed in this chapter. The voice conversion process generally has three parts: spectral modification, residual modification, and prosody modification. Spectral transformation is carried out by transformation function estimated from aligned source and target parameters. The difference of accents, recording environment, or prosody leads to considerable spectral differences in the source and target frames. These differences may result in misaligned frames for the source and the target and the resulting transformation function may be improper due to one-to-many mappings in the aligned frames. The alignment could be improved by using unit-selection framework.

The spectral modification techniques may be grouped into four main categories: vector quantization, ANN and statistical, frequency warping, and speaker interpolation. The quality of the transformed speech using vector quantization is low because of the use of finite number of classes. The ANN and statistical techniques are able to capture the natural transformation function independent of the acoustic unit, but they need a larger set of training data and computation. Frequency warping and speaker interpolation based techniques require less data, but a different transformation function is needed for each acoustic class. Most of the methods for voice conversion use linear transformation for obtaining the mapping function in order to simplify estimation process. This eliminates the second order information from the mapping by over-smoothing, and may result in loss of naturalness in the transformed speech. Transformation using a mapping involving non-linear terms may improve the transformation. Estimation of an efficient transformation function using limited data for obtaining a satisfactory transformed speech quality is still an open challenge to the researchers.

The methods for residue modification may be grouped as source residual, excitation parameterization, excitation conversion, excitation prediction, and excitation selection.

Retaining the source residual provided relatively better speech quality, but residual prediction method provided better similarity of the transformed speech to the target speech. Out of the several methods reported for prosody modification, statistical method appears to be most satisfactory.

Chapter 3

HARMONIC PLUS NOISE MODEL FOR SPEECH MODIFICATION

3.1 Introduction

Several analysis-synthesis platforms, such as source-filter [90], [91], [182], [183], STRAIGHT [137], [184]-[187], wavelet [188], [189], and variants of sinusoidal modeling [5], [20], [190], [191], have been employed for voice conversion. We have selected harmonic plus noise model (HNM) [71], [158], [192], [193], a variant of sinusoidal modeling, as analysis/synthesis platform for voice conversion. It permits a control on the contribution of harmonic and noise bands, and facilitates time and frequency scaling [158], [193]. A description of HNM, as reported in [158], [192]-[197], is given in the next section, and our implementation is described in the subsequent section. The application of HNM for speech modification is presented in Section 3.4. Investigations for evaluating the suitability of HNM as an analysis-synthesis platform for voice conversion are presented in Section 3.5. The last section provides a summary.

3.2 Harmonics plus noise model

In harmonic plus noise model, the speech signal is modeled as the sum of a harmonic signal and a random signal. In voiced segments, the frequency spectrum shows clear peaks at harmonic frequencies up to a certain frequency called maximum voiced frequency (F_m). The synthesized speech using the sum of pitch harmonics up to this frequency is called the harmonic part of the speech. The spectrum above F_m appears to be noisy and the speech corresponding to this part of the spectrum is called the noise part. The harmonic part accounts for the quasi-periodic components of the speech signal while the random or the noise part results due to the non-periodic components (e.g., frication, aspiration, period-to-period variation of the glottal excitation, and modeling errors in the harmonic part) [71].

For speech signal $s(n)$, if $\hat{s}_h(n)$ and $\hat{s}_n(n)$ represent synthesized harmonic and noise parts, then the synthesized speech $\hat{s}(n)$ is given by

$$\hat{s}(n) = \hat{s}_h(n) + \hat{s}_n(n) \tag{3.1}$$

The analysis and synthesis is carried out pitch synchronously, using analysis window of two pitch periods. For voiced segments, the pitch period refers to the local pitch period and for unvoiced segments, the pitch period is taken as 10 ms [192]. In each frame, the harmonic part is a sum of sinusoids with pitch harmonic frequencies with varying magnitudes and phases [198]. Let b_l be the time varying complex amplitudes, F_0 be the local pitch frequency, and F_s be the sampling frequency. The harmonic part may be represented as

$$\hat{s}_h(n) = \sum_{l=-L}^{L} b_l(n) e^{j2\pi l F_0 n / F_s} \tag{3.2}$$

The number of harmonics, L, to be included in each frame, is decided by the local pitch frequency F_0 and the maximum voiced frequency F_m,

$$L = \text{int}(F_m / F_0) \tag{3.3}$$

The values of amplitudes, within each frame, are calculated by interpolating their values at the frame boundaries.

The parameters of the voiced segments (F_0, F_m, and the harmonic amplitudes) are estimated pitch-synchronously. In [158], dynamic programming based pitch estimation, reported earlier by Griffin and Lim [199], is employed to get the first estimate of F_0. For voicing decision, a harmonic approximation error is calculated as

$$\varepsilon = 10 \log \left(\frac{\sum_{k=k_1}^{k_2} \left(|S(k)| - |S_h(k)| \right)^2}{\sum_{k=k_1}^{k_2} |S(k)|^2} \right) \tag{3.4}$$

where $S(k)$ is the DFT of the current frame and $S_h(k)$ is the DFT of the corresponding synthesized harmonic part. The frequency indices k_1 and k_2 correspond to $0.7F_0$ and $4.3F_0$, respectively. If ε is below a threshold (empirically selected value of -15 dB), the frame is declared as voiced else as unvoiced [158].

Estimation of maximum voiced frequency involves identification of the harmonic peaks in the log magnitude spectrum of each voiced frame. Selection of harmonic peaks is based on the peak magnitude, position, its value, and the area between the valleys on its either side, using an algorithm described in [158], [200]. With $f = F / F_s$ being the normalized frequency, let $A(f)$ be the log magnitude spectrum of the signal frame, and f_0 be the normalized pitch frequency. The algorithm can be described as the following

1. Set the starting test voiced frequency as $f_v = 0$.
2. Find the highest peak location in the frequency range $[f_v + 0.5 f_0, \quad f_v + 1.5 f_0]$, and mark it as f_v and its magnitude as A_m.

3. Locate the peaks in the frequency range $[f_v - 0.5f_0, \quad f_v + 0.5f_0]$. For each of the peak locations, calculate the area under the log spectral segment between the valleys on either side of the peak. Let the area for the highest peak be A_0 and average of the areas for all the other peaks be $\bar{A}_m$. The area of the peak next to the highest peak is termed as A_1.
4. Declare f_v as voiced if it is within $\pm 10\%$ of an integral multiple of the pitch frequency f_0 and the harmonic test

   $$\left[\left(A_0 / \bar{A}_m > 2\right) \text{ or } \left(A_m - A_1 > 13\right)\right]$$

   is satisfied.
5. Go to step 2 with the new value of f_v, until the entire band has been covered.

The above search gives a set of L peak locations $f_v(l)$, each declared as voiced (1) or unvoiced (0). This sequence of voiced/unvoiced decisions is smoothened by a 3-point median. The last frequency declared as voiced is labeled as F_m. The value of the pitch frequency f_0 is revised by minimizing the square error between all the voiced frequencies $f_v(l)$ and harmonics of f_0

$$\varepsilon(f_0) = \sum_{l=1}^{L} \left| \left[f_v(l) - l f_0 \right] v(l) \right|^2 \tag{3.5}$$

where $v(l)$ is '1' for voiced frequency and '0' otherwise.

For each analysis window, its center is set as $n = 0$, with the window extending from $n = -N_0$ to N_0, where N_0 is the number of samples in the local pitch period. The synthesized harmonic part as given in (3.2) can be written as

$$\hat{\mathbf{s}}_{\mathrm{h}} = \mathbf{Cb} \tag{3.6}$$

where $\hat{\mathbf{s}}_{\mathrm{h}}$ is the column vector consisting of $2N_0 + 1$ samples of the synthesized harmonic part, $\mathbf{C}$ is $(2N_0 + 1) \times (2L + 1)$ matrix with the (m, n) element given by

$$c_{m,n} = e^{j2\pi(m - N_0 - 1)(n - L - 1) f_0}$$

and $\mathbf{b}$ is the column vector consisting of $2L + 1$ complex harmonic amplitudes. These amplitudes are determined for minimizing the error between the speech signal and the synthesized harmonic part

$$\varepsilon = \sum_{n=-N_0}^{N_0} \left| w(n) \left[s(n) - s_h(n) \right] \right|^2 \tag{3.7}$$

where $w(n)$ is Hamming window. Substituting (3.6) in (3.7), a least-squares solution results in a set of $2L + 1$ simultaneous equations, which can be written as,

$$\mathbf{Rb} = \mathbf{d} \tag{3.8}$$

where $\mathbf{R} = \mathbf{C}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{C}$ and $\mathbf{d} = \mathbf{C}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{s}$. The matrix is $\mathbf{W}$ a $(2N_0+1)\times(2N_0+1)$ diagonal matrix having diagonal elements from Hamming window $w(n)$. Matrix $\mathbf{s}$ is the column vector containing $2N_0+1$ samples of the input speech in the analysis frame. The matrix $\mathbf{R}$ is $(2L+1)\times(2L+1)$ matrix with the (i,k) element given by

$$r_{i,k} = \sum_{n=-N_0}^{N_0} w^2(n) e^{j2\pi(i-k)nf_0}$$

Matrix $\mathbf{d}$ is $(2L+1)\times 1$ matrix with the $(l,1)$ element defined as

$$d_{l,1} = \sum_{n=-N_0}^{N_0} w^2(n)s(n)e^{-j2\pi(l-L-1)nf_0}$$

Here matrix $\mathbf{R}$ is a Toeplitz matrix such that

$$r_{i+p,k+p} = r_{l,k}, \quad \forall i,k,p$$

and Levinson algorithm may be used to solve the set of linear equations. In each of the equations represented in (3.8), only one of the terms (diagonal) dominates, as we can assume negligible interaction among different amplitudes. This is equivalent to assuming the matrix $\mathbf{R}$ as diagonal [158]. This assumption simplifies the calculation of complex harmonic amplitudes and elements of the complex amplitudes vector $\mathbf{b}$ are given by

$$b_l = \frac{\sum_{n=-N_0}^{N_0} w^2(n)s(n)e^{-j2\pi nlf_0}}{\sum_{n=-N_0}^{N_0} w^2(n)}, \qquad -L \le l \le L \tag{3.9}$$

Once, all the necessary parameters for harmonic part of the speech are available, the harmonic part is synthesized using (3.2) which can be rewritten as

$$\hat{s}_h(n) = \sum_{l=1}^{L} a_l(n)\cos(\phi_l(n)) \tag{3.10}$$

where $a_l(n)$ and $\phi_l(n)$ are the magnitudes and phases of the harmonics at each sample within the frame, which may be estimated by interpolation of the magnitudes and phases obtained during the analysis. The phase $\phi_l(n)$ can be treated as a sum of two terms, a system phase for a specific frame, and a linear phase which changes linearly with frequency and sample position within the frame. Stylianou [196] investigated three interpolation methods: step, linear, and quadratic. In the first case, the resynthesis is carried out using (3.2). The complex amplitudes for each harmonic are retained constant during the whole frame and phases are linearly varied without using the slope of the analysis phases. In the second case, the resynthesis is carried out using (3.2), but the complex amplitudes and phases are linearly interpolated. The third method uses (3.10) with $\phi_l(n)$ taking care of both system and linear

phases. The harmonic magnitudes are interpolated using a quadratic function and phases are varied linearly from the initial system phases. Relatively smaller modeling errors were reported in the second and third case [196].

The noise part is obtained by subtracting the synthesized harmonic part from the original speech. The parameters of the noise part are obtained by 10-15 order linear predictive (LP) analysis [192]. The analysis window is twice the local pitch period with shift of one pitch period if the current frame is voiced. If the current frame is unvoiced, the window length is taken as 20 ms with shift of 10 ms. The other parameter of the noise part is short-time energy, estimated over 2 ms segments.

Synthesized noise part is obtained by applying a unit variance and zero mean white Gaussian noise $r(n)$ to an all-pole filter, and multiplying the filter output by gain $G(n)$,

$$\hat{s}_n(n) = G(n)\big(h(n) * r(n)\big) \tag{3.11}$$

where $h(n)$ is the all-pole filter defined by the LPC coefficients obtained during analysis of the noise part. $G(n)$ is obtained by interpolation of the square root of the short-time energy [158], [192], [196].

For time or pitch scaling, the synthesis time instants are estimated from the analysis time instants using the given scaling factor. The HNM parameters at analysis instants are linearly interpolated for estimating the parameters at the synthesis instants. These parameters are used for generating the speech at each synthesis time instant. The main advantage of this approach is that it eliminates most of the artifacts associated with the pitch-synchronous overlap-and-add (PSOLA) method [71], [201].

The quality of the synthesized speech may be affected by modification of the parameters. Text-to-speech (TTS) and voice conversion involve modification of the parameters. As the speech units may not be obtained from consecutive frames, there concatenation may produce distorted quality due to phase mismatches, which may be caused by mismatch in linear phase or system phase [158], [195], [202]. The linear phase mismatch produces garbled speech in voiced segments, but it does not affect the unvoiced segments. System phase mismatch is introduced by different distributions of the system phase around concatenation. It is responsible for induction of noise between harmonic peaks and is perceived as continuous background (mostly during unvoiced sounds). Linear phase mismatch is perceptually more important than system phase mismatch.

3.3 HNM implementation

In our implementation, some of the analysis and synthesis blocks of the model described in the previous section, have been modified for reducing the computational complexity and improving the flexibility in modification of the parameters for voice conversion. The first

modification is in the pitch-synchronous analysis of speech. We use glottal closure instants (GCI) based pitch-synchronous processing. It was reported by Wang *et al*. [203] that the position of the maximum excitation is crucial for accurate estimation of parameters, particularly for low and high pitch periods. The methods used for pitch frequency estimation and voiced/unvoiced decision are different from those in [158]. For the analysis and synthesis of the noise part, use of LSFs and MFCCs has been explored, because of suitability of these parameters for modification and interpolation. The synthesized harmonic part and synthesized noise part are generated separately, using the parameters obtained during the analysis of the input speech signal. The synthesized speech in each frame is given by the temporal summation of these two parts.

3.3.1 Analysis of harmonic part

A block diagram for the HNM based analysis is shown in Fig. 3.1. The speech signal is applied to the voicing detector, which declares the frames either as voiced or as unvoiced. For pitch-synchronous analysis and synthesis, we have used Childers and Hu's algorithm [88], [204] for detecting GCIs and the pitch frequency. Voiced segments are further analyzed for obtaining maximum voiced frequency, harmonic magnitudes, and harmonic phases. As the coefficients used in processing have been earlier empirically established for 10 kHz sampling, the speech signal was converted from 16 kHz to 10 kHz sampling frequency for processing by this algorithm and the detected GCIs locations were converted to those corresponding to the original sampling rate. An examination of the results of pitch tracking did not show any errors due to sampling rate conversion.

Voiced / unvoiced decision: The speech signal is first divided into frames of duration 25 ms each with 50% overlap and a decision is taken on the nature of the frame, i.e. voiced or unvoiced using the Childers and Hu algorithm [88], [204]. This method uses first reflection coefficient defined as

$$r_1 = \frac{R_{ss}(1)}{R_{ss}(0)} = \frac{\frac{1}{N}\sum_{n=0}^{N-1} s(n)s(n+1)}{\frac{1}{N}\sum_{n=0}^{N-1} s(n)s(n)} \tag{3.12}$$

where N = frame length. For unvoiced segments, this ratio is generally very low. Decision of voicing is taken according to values of reflection coefficient and the energy of the prediction error (LPC residue) e_p and empirically selected thresholds. Let T_e be the empirically chosen threshold for e_p. If $e_p > 2T_e$ and $r_1 > 0.2$ for the current frame, the frame is declared as voiced. If the previous frame is voiced and $e_p > T_e$ along with $r_1 > 0.3$ for the current frame,

the frame is declared as voiced. If none of these conditions holds true, the frame is declared as unvoiced. The voicing decision $v(i)$ for the current frame i can be written as

$$v(i) = \left(\left(e_p > 2T_e \right) \wedge \left(r_1 > 0.2 \right) \right) \vee \left(v(i-1) \wedge \left(e_p > T_e \right) \wedge \left(r_1 > 0.3 \right) \right) \tag{3.13}$$

The voiced and unvoiced frames are denoted by '1' and '0', respectively. It is further assumed that voiced or unvoiced segments last for at least two consecutive frames. The sequence of voiced/unvoiced decision is processed to eliminate abrupt transitions using a moving five-point window. If the central value in the window is the complement of all the other values, it is complemented so that they all become the same. Thus the patterns 11011 and 00100 are replaced by 11111 and 00000, respectively.

Detection of glottal closure instants: Childers and Hu's algorithm [88], [204], [205] has been used for the detection of GCI. The LPC prediction error $e(n)$ is integrated and smoothened by using the following filters.

$$H_1(z) = \frac{1 - z^{-1}}{1 - 0.99z^{-1}} \text{ and } H_2(z) = \frac{1 - z^{-1}}{1 - 1.6z^{-1} + 0.63z^{-2}}$$

The filtered output of each filter is reversed and again applied to the same filter to maintain zero phase distortion. In the smoothened signal $\hat{e}_{LP}(n)$, the most negative peak is located. The speech signal in its neighborhood (15 samples before and 30 samples after) is cross-correlated with $\hat{e}_{LP}(n)$, and the positive peaks in the cross-correlation are taken as GCI positions. It may be noted that there may be some duplication of peak locations in the consecutive frames and these are corrected during the process of estimation of pitch periods as described later.

Estimation of pitch periods: Estimation of pitch periods is a two-pass process. In the first pass, these are estimated as the intervals between successive GCI locations. In the second pass, these are refined using the peak picking algorithm [88], [204], [205]. First, a cepstrum-like function is calculated by taking FFT (N = 256) of the magnitude spectrum of $\hat{e}_{LP}$ as

$$c_e(n) = \text{IFFT}(\left|\text{FFT}(\hat{e}_{LP}(n))\right|) \tag{3.14}$$

Let the index of the maximum peak in $c_e(n)$ in the range $[24, \quad N-1]$ be m. For this peak, let the maximum value of $c_e(n)$ in the range $[24, \quad m-24]$ be at k. If $c_e(k) > 0.7c_e(m)$, k is taken as the pitch period otherwise m is retained. These steps are repeated for every voiced frame and a smooth pitch contour is obtained by applying a 3-point median filter on the sequence of pitch periods. If pitch periods are not found to be equal in the two passes, pitch periods and the corresponding GCIs are corrected by considering pitch doubling or halving. This also corrects duplication of some of detected GCIs in consecutive frames.

Harmonic magnitudes and phases: Estimation of complex harmonic amplitudes is carried out by a pitch-synchronous method, fixing the analysis window at each GCI spanning from the previous GCI to the next GCI and using (3.9). In case of a voiced-to-unvoiced (or unvoiced-to-voiced) transition, the window length is set by assuming a dummy GCI in the subsequent (or preceding) frame at a location corresponding to the current pitch period. The complex amplitudes are converted into harmonic magnitudes and phases.

3.3.2 Analysis of noise part

In voiced frames, the noise part is obtained by subtracting the synthesized harmonic part from the original speech signal. For unvoiced frames, the whole frame is treated as the noise part. The different steps involved for the analysis are shown in Fig. 3.1. We have experimented with two methods for modeling of the envelope of the noise part. In the first method, the spectral envelope of the noise part is modeled as autoregressive filter and LPC coefficients are calculated. Because of difficulties in modifying the LPC coefficients for voice conversion, they are converted to LSFs. In the second method, noise part is converted to MFCCs. For voiced frames, the analysis window is the same as that for the harmonic part analysis. For unvoiced frames, it is taken as 20 ms with 50% overlap. The temporal variation of the energy in the harmonic and noise bands is important for the perception of stops, fricatives, aspirates, and other transitional speech units. Hence, the energy of the noise part is estimated over 2 ms segments.

3.3.3 Synthesis of harmonic part

A block diagram of the pitch-synchronous HNM synthesis is shown in Fig. 3.2. The voiced speech segments can be assumed to be quasi-periodic, and hence the parameters at each sampling instant within the frame are calculated by interpolating the parameters available at frame boundaries.

The harmonic part of the signal in a voiced frame is calculated as [196]

$$\hat{s}_h(n) = \sum_{l=1}^{L} a_{l,i}(n)\cos(\phi_{l,i}(n)), \qquad 0 \le n \le N_0\text{-}1 \tag{3.15}$$

where $a_{l,i}(n)$ and $\phi_{l,i}(n)$ are interpolated magnitudes and phases for frame i and harmonic l, respectively, for sample n in frame i. The local pitch period N_0 in frame i is the interval between two successive GCI locations, i.e. as $N_0 = \mathrm{GCI}(i+1) - \mathrm{GCI}(i)$.

The frame index i can also be treated as GCI i, as the analysis window is centered on each GCI. For each frame i, harmonic magnitude within the frame for harmonic l is obtained by linear interpolation of its values at GCI locations i and i+1, and it is given by

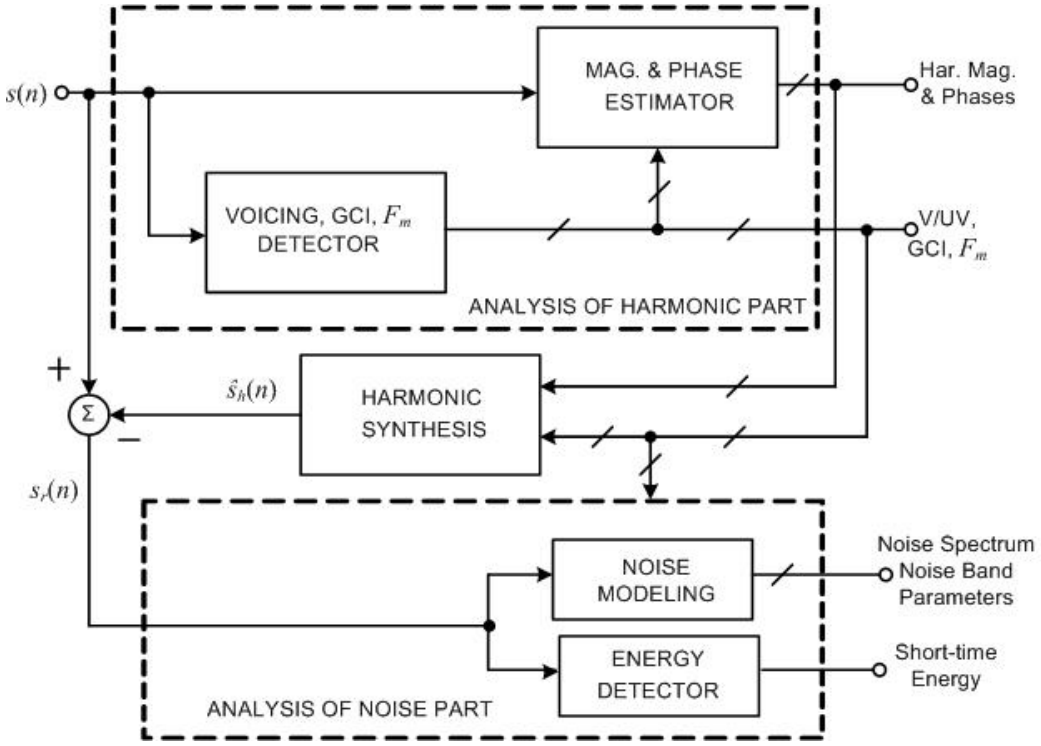

Fig. 3.1 HNM based speech analysis.

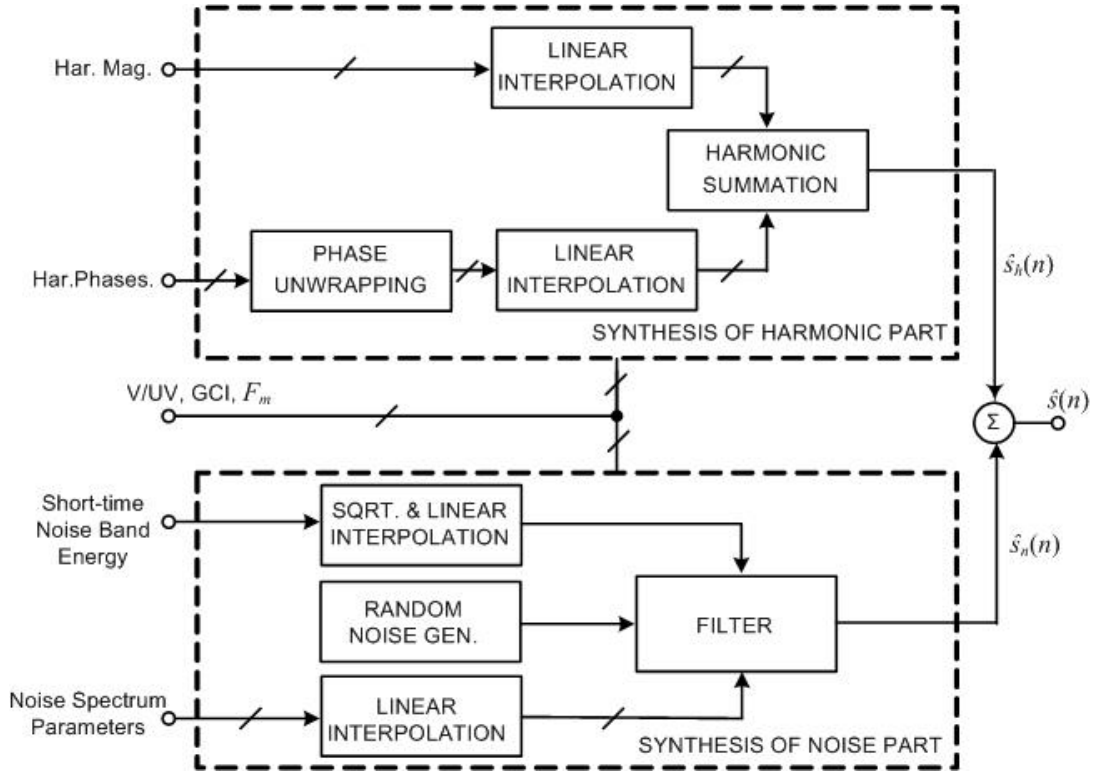

Fig. 3.2 HNM based speech synthesis. The parameters V/UV, GCI, and F_m are used in linear interpolation, harmonic summation, and filtering.

$$a_{l,i}(n) = a_{l,i} + \frac{a_{l,i+1} - a_{l,i}}{N_0} n, \quad 1 \le l \le L, \quad 0 \le n \le N_0\text{-}1 \tag{3.16}$$

The instantaneous phase is a function of time and frequency. As the phases available at frame boundaries obtained from analysis are modulo 2π, these need to be unwrapped before interpolation. Let $f_{0,i}$ be the normalized pitch frequency and $\phi_{l,i}$ be the estimated phase for harmonic l in frame i. The phase for frame $i+1$ can be predicted from the phase of the previous frame as

$$\breve{\phi}_{l,i+1} = \phi_{l,i} + 2\pi l f_{0,i} N_0, \quad 1 \le l \le L \tag{3.17}$$

The actual value of phase obtained from the analysis is modified by adding an integer multiple of 2π to the value of phase for frame $i+1$ and harmonic l so that it is closest to the predicted value

$$\hat{\phi}_{l,i+1} = \phi_{l,i+1} + 2\pi M_l, \quad 1 \le l \le L \tag{3.18}$$

where M_l is obtained as

$$M_l = \left\langle \frac{\breve{\phi}_{l,i+1} - \phi_{l,i+1}}{2\pi} \right\rangle, \quad 1 \le l \le L \tag{3.19}$$

Now the phase for the frame i is estimated by

$$\phi_{l,i}(n) = \phi_{l,i} + \frac{\hat{\phi}_{l,i+1} - \phi_{l,i}}{N_0} n, \quad 1 \le l \le L, \quad 0 \le n \le N_0\text{-}1 \tag{3.20}$$

If there is a transition from voiced frame to unvoiced frame or vice-versa, the harmonic magnitudes of the frame spanning into unvoiced part are set to zero and phases are calculated by extrapolating the phases in (3.20).

3.3.4. Synthesis of noise part

The scheme for synthesizing noise part is shown in Fig. 3.2. The length of synthesis window is taken as 2 ms without any overlap. The noise parameters obtained from analysis are interpolated to get their values at the window locations. For LSF-based synthesis, LSFs are converted to LPC coefficients. A unit-variance zero-mean Gaussian noise is applied as input to a unit gain filter defined by LPC coefficients. For MFCC-based synthesis, MFCCs are converted to spectrum, as described later in Subsection 3.4.2. The spectrum is multiplied by DFT of a unit-variance Gaussian noise and the noise part is obtained by taking its inverse DFT and retaining the initial samples corresponding to the window length. In both synthesis methods, the output is multiplied by a gain, obtained by interpolating the square root of the short-time energy for each frame.

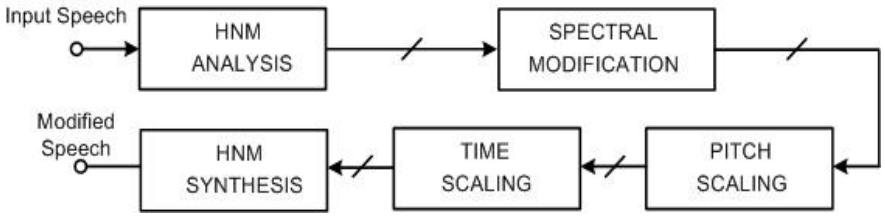

Fig. 3.3 HNM based speech modification.

3.4 Speech modifications using HNM

In voice conversion, three types of modifications are required: pitch, time, and spectral envelope. Pitch modification is carried out by pitch scaling to match the range of the pitch in the source speech to that in the target speech. Time modification is carried out by time scaling to match the duration of the source speech to that of the target. Pitch-scaling involves a modification of the pitch contour without disturbing the duration of the speech signal. Time-scaling modifies the duration of the speech signal, keeping the pitch contour intact. Spectral envelope modification involves modifying the spectral characteristics of the source speech to match them with the target speech. The spectral envelope is contributed by the shape of the excitation pulse as well as the vocal tract filter. However the vocal tract filter is considered to be the main contributor and hence the spectral envelope modification is also known as vocal tract transformation.

Fig. 3.3 shows the basic scheme used for HNM-based modification of speech. The parameters of the speech signal are obtained from HNM analysis and modified for the spectral modification. The synthesis axis is prepared according to the given pitch and time scaling factors. HNM parameters are estimated at the instants on the synthesis axis using interpolation. The modified parameters on the synthesis axis are used for synthesizing the speech output.

In the previous section, we have described two techniques for synthesizing the noise part. These require a set of parameters which are different from those in the harmonic part and hence need to be handled separately during voice conversion. To address this problem, it is proposed to synthesize the speech in the voiced frame using harmonically sampled magnitudes of the spectrum, and associating the magnitudes below F_m with the estimated phases and those above it with random phases, thus synthesizing the noise part without estimating it separately. Therefore, we have three variants of HNM. All three use the same

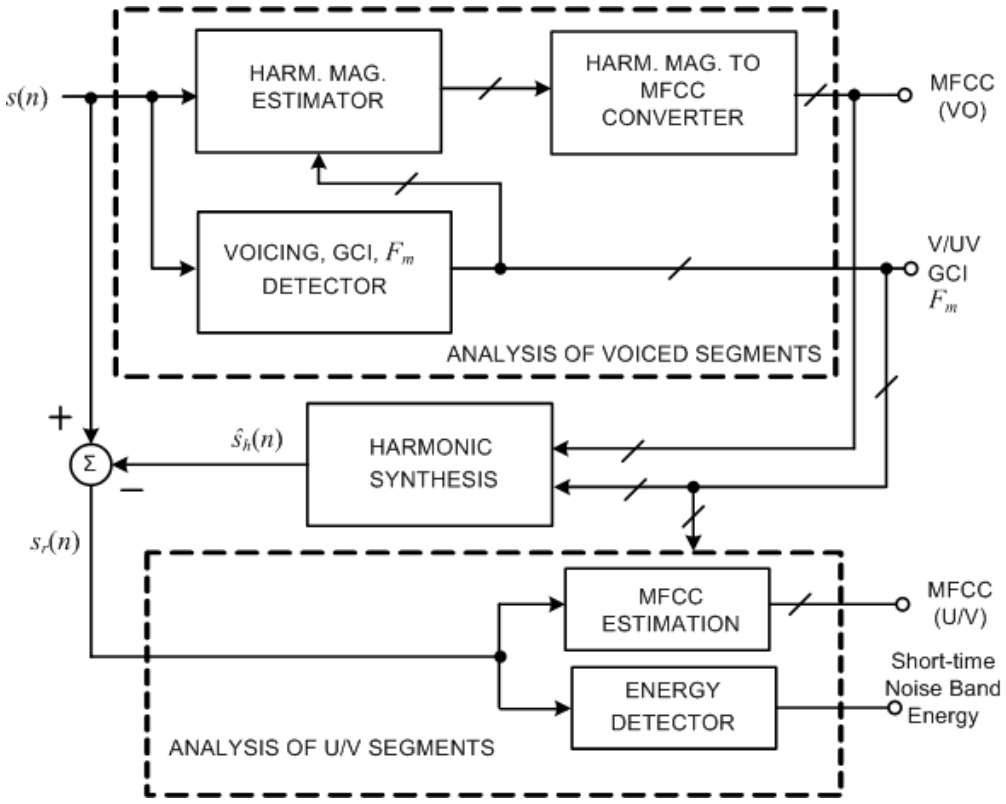

Fig. 3.4 HNM-3 based speech analysis

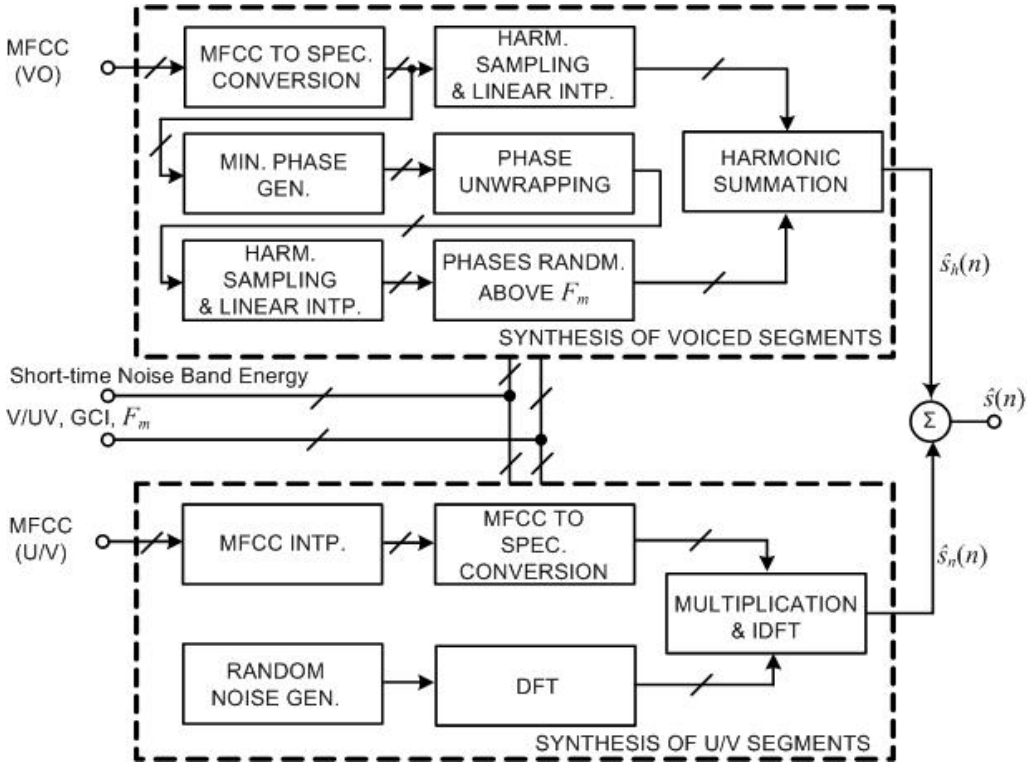

Fig. 3.5 HNM-3 based speech synthesis. The parameters V/UV, GCI, and F_m are used in linear interpolation, harmonic summation, and filtering. The short-time energy is used for energy adjustment of noise part in harmonic summation and noise synthesis in unvoiced frames.

method for the harmonic part of the voiced frames, and differ only in modeling of the noise part. HNM-1 uses LSFs for representing the noise part and HNM-2 uses MFCCs. In HNM-3, no separate synthesis of the noise part in voiced frames is carried out, except that the contribution of the harmonics above F_m is estimated by giving random values to their phases. The noise in the unvoiced frames is modeled using MFCCs.

The analysis and synthesis using HNM-3 with respect to speech modification is shown in Fig. 3.4 and Fig. 3.5. During analysis, all harmonic magnitudes estimated upto $F_s/2$ are converted to discrete MFCCs using the technique described in Subsection 3.4.1. The noise part in the voiced frames is obtained by subtracting the harmonic part (synthesized upto 4 kHz) for computing short-time energy using 2 ms segments. The noise part in unvoiced frames is converted to MFCCs using the technique described in Subsection 3.4.3. For synthesis of the harmonic part, the MFCCs are converted to magnitude spectrum and phases are estimated using minimum phase assumption, described in Subsection 3.4.2. After phase unwrapping and linear interpolation, the phases of the harmonics above 4 kHz are assigned values using a uniform random distribution and harmonic summation is carried out. The contribution of the sinusoids above 4 kHz is adjusted according to the short-time energy of the noise part. The process for synthesis of the noise part in unvoiced frames is the same as used for HNM-2 and described earlier in Subsection 3.3.4.

3.4.1 Spectral magnitude function for voiced frames

HNM analysis provides harmonic magnitudes at the multiples of the pitch frequency. For pitch modification, the harmonic magnitudes are needed to be estimated at a different set of harmonic frequencies. In [206], a technique is suggested for estimating a continuous magnitude spectrum as a function of frequency from the discrete harmonic magnitudes. It passes through all the given harmonics magnitudes and has a smooth variation in between the harmonic values, and therefore it can be used for interpolating the magnitudes at the desired frequencies. The parameters of the spectral magnitude function are similar to MFCCs [95], if the harmonic frequencies are warped in accordance to the mel scale [143]. As the parameters are obtained from discrete set of frequencies, they are also known as discrete MFCCs [145]. Because the spectrum is relatively accurately defined at harmonic frequencies, the discrete MFCCs provide a better fit to the spectral envelope [20], [143], [144], [206].

The conversion of frequency F to mel scale [207] is given by

$$\text{mel}(F) = 2595 \log\left(1 + \frac{F}{700}\right) \tag{3.21}$$

Let $\mathbf{a} = \begin{bmatrix} a_0 & a_1 & a_2 & \cdots & a_L \end{bmatrix}^{\mathrm{T}}$ be the column vector containing harmonic magnitudes at the set $\begin{bmatrix} 0 & F_1 & F_2 & \cdots & F_L \end{bmatrix}^{\mathrm{T}}$ of harmonic frequencies ($F_l = lF_0$) for a frame i. It has been shown in [206] that the discrete MFCCs can be estimated as

$$\mathbf{c} = \left(\mathbf{M}^{\mathrm{T}}\mathbf{W}\mathbf{M} + \lambda\mathbf{R}\right)^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{W}\log(\mathbf{a}) \tag{3.22}$$

where $\mathbf{W}$ is a diagonal matrix having diagonal elements as Hamming window samples. As suggested in [206], the value of regularization parameter λ is taken as 5×10^{-4} to avoid ill conditioning during the matrix inversion. The matrix $\mathbf{R}$ is given by

$$\mathrm{Diag}(\mathbf{R}) = 8\pi^2\begin{bmatrix} 0 & 1^2 & 2^2 & \cdots & p^2 \end{bmatrix} \tag{3.23}$$

and $\mathbf{M}$ is a $(L+1)\times(p+1)$ matrix with element (l,m) given by

$$e_{l,m} = 2\cos(2\pi(m-1)\ \mathrm{mel}(F_{l-1})/\mathrm{mel}(F_s)), \qquad 1\le l\le L+1,\ 1\le m\le p+1 \tag{3.24}$$

The spectral log magnitude function $S(F)$ is reconstructed from *p* discrete MFCC coefficients using the following relation [206]

$$S(F) = c_0 + 2\sum_{m=1}^{p} c_m \cos(2\pi m\ \mathrm{mel}(F)/\mathrm{mel}(F_s)) \tag{3.25}$$

The log magnitude function is evaluated at *K*/2 uniformly spaced frequency samples over [0, F_s/2]. These samples are used to get the discrete frequency spectrum as the following

$$\tilde{S}(k) = \begin{cases} \exp(S(kF_s/K)), & 0\le k\le K/2 \\ \tilde{S}(K-k), & K/2+1\le k\le K-1 \end{cases} \tag{3.26}$$

with *K* = 1024. Our investigations were carried out using different values of *p*. Based on the analysis-synthesis results, *p* = 20 was used for voice conversion.

3.4.2 Phase reconstruction for voiced frames

In the transformation process of the harmonics, only the magnitude information is retained and phase information is lost. Many methods are available for associating a phase function with a given magnitude function for use in analysis-synthesis systems [208]-[212]. We have used a non-iterative estimation of minimum phase function [213], [214] followed by an iterative method for phase reconstruction [212]. The magnitude spectrum $\tilde{S}(k)$ is used to calculate the real cepstrum

$$\tilde{c}(m) = \mathrm{IDFT}\left[\log\left(\tilde{S}(k)\right)\right] \tag{3.27}$$

We then calculate a complex cepstrum using a minimum phase function assumption [213], [214]

$$\hat{s}(m)=\begin{cases}\tilde{c}(m), & m=0,K/2\\ 2\tilde{c}(m), & 1\le m\le K/2-1\\ 0, & K/2+1\le m\le K-1\end{cases} \tag{3.28}$$

From the complex spectrum corresponding to $\hat{s}(m)$, we calculate the phase spectrum $\phi_0(k)$, as the initial estimation of the phase function for the iterative method. The magnitude spectrum $\tilde{S}(k)$ and $\phi_0(k)$ are used through an iterative process to estimate phase spectrum $\phi(k)$. After *j*th iteration, we get $V_j(k)=\tilde{S}(k)\angle\varphi_j(k)$

$$v_j=\text{IDFT}\left[V_j(k)\right] \tag{3.29}$$

The next iteration of the sequence is calculated by imposing the following condition [212]

$$v_{j+1}(n)=\begin{cases}s(0), & n=0\\ v_j(n), & 1\le n\le K/2\\ 0, & K/2+1\le n\le K-1\end{cases} \tag{3.30}$$

where $s(0)$ is known from the previous placing of the synthesis window, as the windows have one pitch period overlap. The complex spectrum is calculated as

$$V_{j+1}(k)=\text{DFT}\left[v_{j+1}(n)\right]$$

The corresponding phase spectrum $\phi_{j+1}(k)$ is taken as the revised phase estimate and it is combined with the original magnitude spectrum $\tilde{S}(k)$ for the next iteration. The iteration is halted if the mean square error between $\left|V_{j+1}(k)\right|$ and $\tilde{S}(k)$ is below a set threshold.

The complex spectrum obtained from the original magnitude spectrum and the estimated phase spectrum is interpolated at harmonic frequencies to get a_l and ϕ_l. The estimated harmonic phases are unwrapped along the frequency and time axes. For unwrapping along frequency axis, a process described in [215] has been used. It makes the instantaneous frequency continuous around each harmonic by adding multiples of 2π to the phases for keeping the variation of phase angles with frequency index as smooth as possible. Let $\phi_{l,i}$ be the phase in frame i for harmonic l. The unwrapped phase for harmonic l is obtained as

$$\breve{\phi}_{l,i}=\phi_{l,i}+2\pi M_l \tag{3.31}$$

with the integer M_l such that the changes in the unwrapped phases of harmonic l -1 and harmonic *l* with respect to frequency index become approximately equal, i.e.,

$$\breve{\phi}_{l-1,i}-\breve{\phi}_{l-2,i}\approx\breve{\phi}_{l,i}-\breve{\phi}_{l-1,i}$$

This is obtained by selecting

$$M_l = \text{int}\left(\frac{(\breve{\phi}_{l-1,i} - \breve{\phi}_{l-2,i}) - (\phi_{l,i} - \breve{\phi}_{l-1,i})}{2\pi} \right) \tag{3.32}$$

Now these unwrapped phases can be interpolated at any set of frequencies to estimate the harmonic phases at different pitch values. Unwrapping along time axis has already been described in Subsection 3.3.3.

3.4.3 Estimation of MFCCs for unvoiced frames

For estimating the MFCCs [98], [205], [206], [216], [217] of an unvoiced frame, the power spectrum of the frame is obtained from its DFT

$$P(k) = |X(k)|^2, \qquad 0 \le k \le K-1 \tag{3.33}$$

where K is the DFT size (1024). These are converted to band energies of q bands with equal bandwidth on the mel scale,

$$E_i = \sum_{k=0}^{\frac{K}{2}-1} \beta_i(k) P(k), \qquad 1 \le i \le q \tag{3.34}$$

where β_i is a weighted triangular function associated with the band i on mel scale [206] with the following constraint.

$$\sum_{k=0}^{\frac{K}{2}-1} \beta_i(k) = 1 \tag{3.35}$$

Now MFCCs are calculated as DCT of the log of E_i as given by

$$c_m = \sum_{i=1}^{q} \log_{10}(E_i) \cos(j \frac{\pi}{m}(i+0.5)), \qquad 0 \le m \le p \tag{3.36}$$

The magnitude spectrum can be estimated from the MFCCs by padding MFCCs with zeros to the dimensionality of the filter, inverse DCT, exponentiation, and inverse mel-scale weighting. The inverse mel-scale weighting was applied using the pseudo-inverse of the mel-weights matrix as described in [216]. The investigations for voice conversion were carried out using q = 40 and p = 20.

3.4.4 Pitch scaling

The purpose of pitch scaling is to change the pitch contour while preserving the short-time-spectral envelope. It does not alter the locations and bandwidths of the formants, or their variation with respect to time [215], [218]. An example of pitch contour in pitch scaling is shown in Fig. 3.6. Let the relation between the original and desired pitch contours be given as

$$F_0'(t') = \frac{1}{\alpha(t)} F_0(t) \tag{3.37}$$

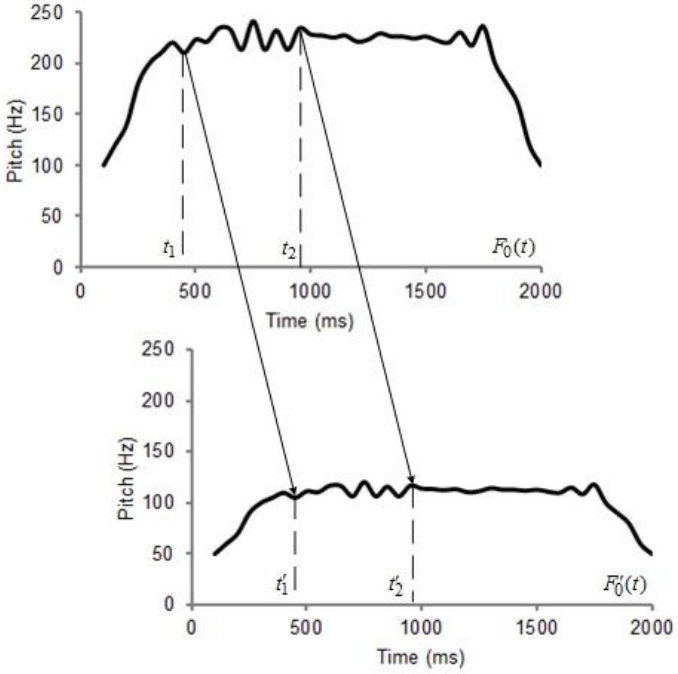

Fig. 3.6 An example of pitch contour in pitch scaled speech signal (2:1).

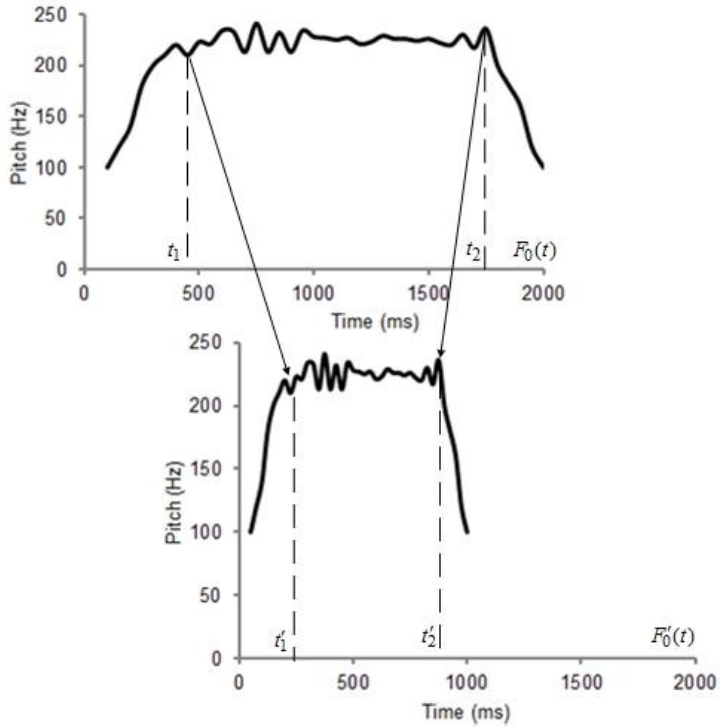

Fig. 3.7 An example of pitch contour in time scaled speech signal (2:1).

where t and t' are the time instants on the analysis and synthesis axes, respectively. For voiced segments of speech, signal parameters are estimated using an analysis window of two pitch periods and one pitch-period shifting. The first GCI in each voiced segment is not disturbed during pitch scaling. The subsequent GCIs are obtained by adding the pitch period obtained by sampling the modified pitch contour at the position of the previous GCI. Harmonic magnitudes are estimated at the GCI locations on the synthesis axis by interpolation of the magnitudes available at analysis instants. Harmonic phases are unwrapped along frequency before interpolation. The harmonic magnitudes obtained after interpolation are used for calculating the spectral magnitude function which is sampled at new pitch harmonics. The unvoiced segments are analyzed using 20 ms window with 50% shift, without any modification for pitch-scaling. Synthesis is carried out using modified HNM parameters for obtaining the pitch modified speech signal using the scheme shown in the Fig. 3.5.

3.4.5 Time scaling

Time scaling is used to change the rate of articulation of the speech maintaining the shape of the pitch contour unchanged. This requires the pitch contour to be stretched or compressed in time and making the formant structure to evolve at a slower or faster rate than the rate of the input speech [215]. An example of the pitch variation during time scaling is shown in Fig. 3.7. Let $F_0'(t')$ be the time-scaled version of the pitch contour $F_0(t)$. The events taking place at an instant t' on the synthesis axis and the corresponding instant t on the analysis axis are related by the relation $t = D^{-1}(t')$. The amount of scaling of a time segment on analysis axis depends upon the warping function $D(t)$ given as

$$D(t) = \int_0^t \beta(\tau)\mathrm{d}\tau \tag{3.38}$$

where $\beta(t)$ is time-modification rate. The value of $\beta(t) > 1$ corresponds to slowing down the rate of articulation by means of a timescale expansion, and the value of $\beta(t) < 1$ corresponds to speeding up the rate of articulation by timescale compression.

The first step for time scaling is similar to pitch scaling. The input speech is analyzed for estimating the speech parameters using an analysis window of two pitch period for voiced segments and 20 ms for unvoiced segments with 50% shifting in both. According to the pitch-scaling factor, GCI locations are modified for obtaining instances on the synthesis axis. As the unvoiced segments are not modified, the first GCI locations on both analysis and synthesis axes are the same. For obtaining the next GCI, the modified pitch contour is sampled at this instant on the synthesis axis. The time period corresponding to the sampled pitch frequency is added to the first GCI and the process is repeated for estimating the remaining GCI. The HNM parameters of the speech are estimated at these GCIs from the known parameters at the

analysis axis according to the given time-scaling factor. The process of interpolation is the same as used for pitch-scaling. The modified parameters are used for the synthesis of time-scaled version of the input speech.

Voice conversion generally involves pitch as well as time scaling. First, the synthesis parameters are obtained for pitch scaling. These parameters are then used as the input parameters for obtaining the synthesis parameters for time scaling. The resulting parameters are used for synthesis of the output speech, as shown in Fig. 3.5.

3.5 Investigations using HNM

A set of six investigations related to some of the issues important for voice conversion were carried out. The first investigation compared HNM variants, HNM-1, HNM-2, and HNM-3, for analysis and synthesis. As described earlier in Section 3.3.2, the three differ in the way the noise part is handled. The effect of maximum voiced frequency F_m separating the harmonic and the noise bands in voiced frames is examined in the second investigation. Errors in estimating the GCIs used for pitch-synchronous analysis and synthesis may lead to degradation of the synthesized speech quality. The effect of these errors is examined in the third investigation. Effect of input SNR is examined in the fourth investigation. During voice conversion, the source magnitude spectrum is modified and its relationship with the source phase spectrum is lost. Therefore the phase spectrum is estimated from the magnitude spectrum. The effect of the estimated phase on the synthesized speech quality is examined in the fifth investigation. The sixth investigation is conducted to study the ability of HNM based analysis-synthesis for pitch and time scaling.

These investigations are carried out using three Hindi sentences from six speakers (3 male, 3 female, age: 20–23 years, mother tongue: Hindi). The speech material was recorded with 16 kHz sampling and 16-bit quantization in an acoustically treated room using Sony ICD-PX820 audio recorder. For comparison of the quality of the synthesized speech in these investigations, objective evaluation based on PESQ-MOS, as described in Appendix C, is used.

3.5.1 Investigation I: Effect of HNM variants

This investigation compared the quality of synthesized speech of three HNM variants: HNM-1, HNM-2, and HNM-3. In HNM-3, the maximum voiced frequency was fixed at $F_s/2$ during analysis and at 4 kHz during synthesis. The noise part in the voiced frames is obtained by subtracting the harmonic part synthesized upto 4 kHz for estimating the short-time energy at 2 ms segments. A comparison was carried out visually using spectrograms, by objective measure of PESQ-MOS, and through listening.

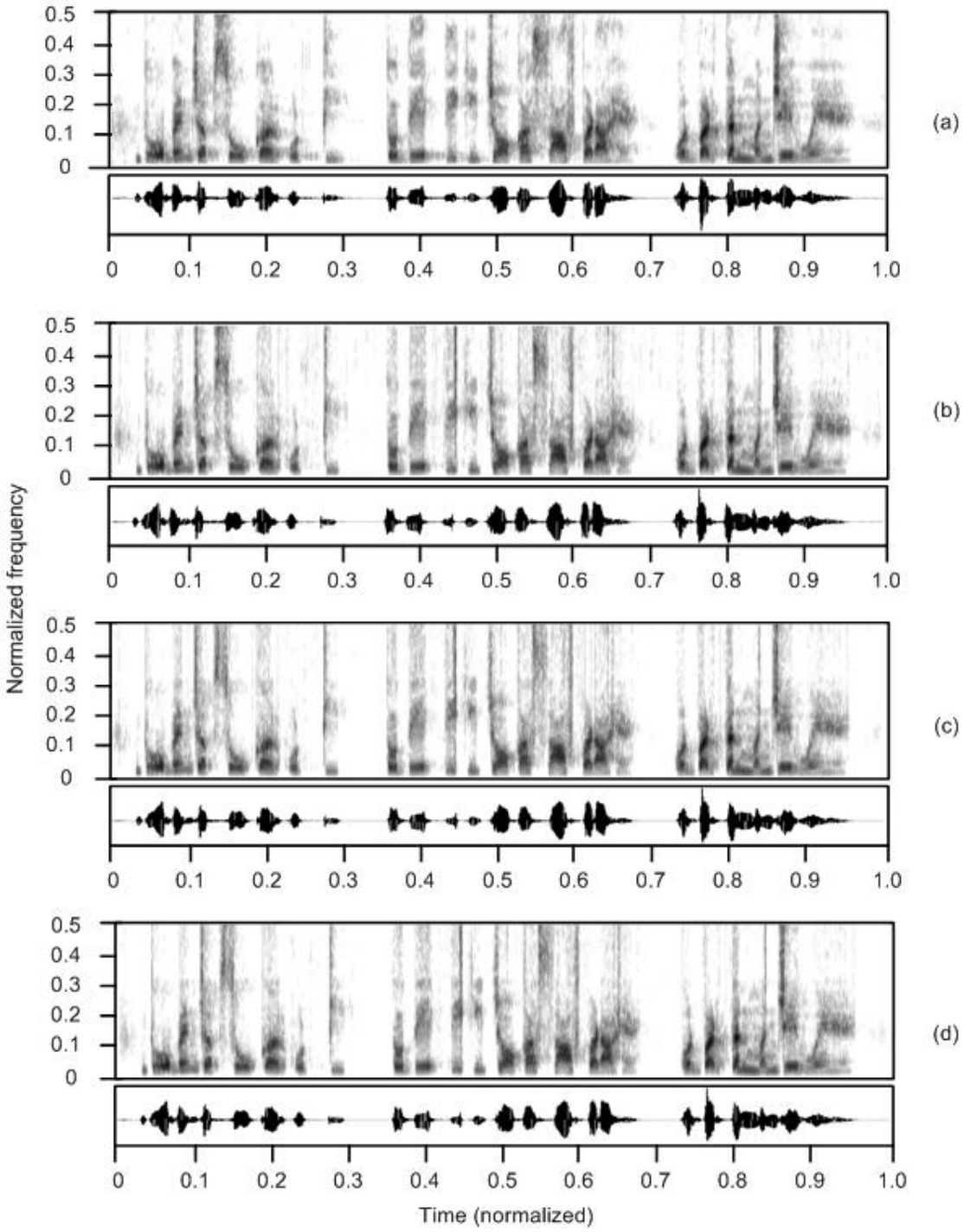

Fig. 3.8 Investigation I: Effect of HNM variants. Spectrograms of the Hindi utterance (F_s = 16 kHz, duration = 7.37 s) /d̪ʰoːbiːn̪ dʒəb soːkər ʊt̪ʰt̪i t̪oː d̪ekʰt̪i ki tʃoːkaː saːpʰ pəd̪aː həi ɔːr bərt̪ən mən̪dʒɛː hʊɛː həiːn̪/ spoken by a female speaker (F1). a) recorded, b) synth. using HNM-1, c) synth. using HNM-2, and d) synth. using HNM-3.

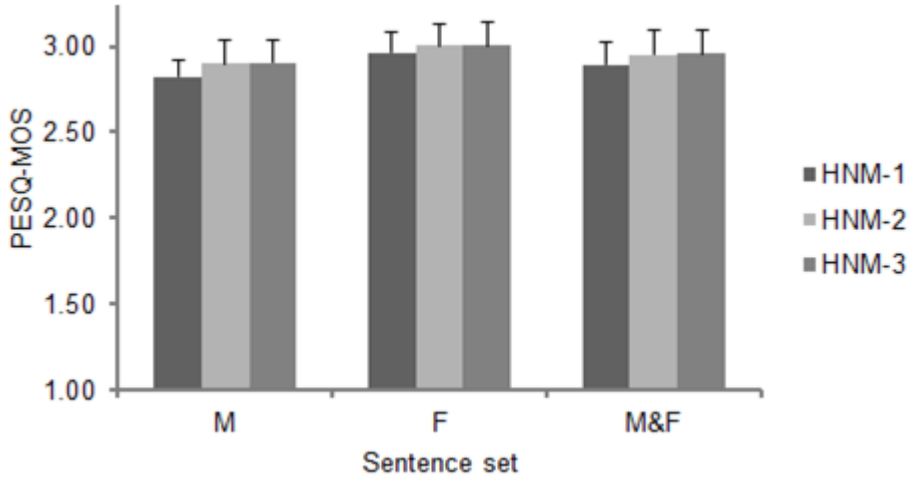

Fig. 3.9 Investigation I: PESQ-MOS test scores of the synthesized utterances (with recorded as reference), averaged over the sets of utterances from male (M), female (F), and male and female (M&F) speakers, for HNM-1, HNM-2, and HNM-3 based analysis-synthesis.

Fig. 3.8 shows spectrograms for one of the utterances from a female speaker (F1). The spectrograms derived from the speech signals synthesized by all three HNM variants are smooth and very near to the spectrogram of the corresponding recorded signal. Similar results were observed in the spectrogram for all the utterances. The means along with the standard deviations of PESQ-MOS test scores are listed in Table A.1 in Appendix A and plotted in Fig. 3.9. The scores for female speech were slightly higher than the corresponding scores for the male speech. Although the score for HNM-2 and HNM-3 are higher than those for HNM-1, the differences are not statistically significant. Also, no difference between the quality of the synthesized speech using the three variants could be heard. Based on these results, we have chosen HNM-3 based implementation for further use because of ease of noise part modeling during voice conversion.

3.5.2 Investigation II: Effect of maximum voiced frequency on synthesized speech

In HNM, the maximum voiced frequency F_m separates the harmonic band from the noise band in voiced frames. It varies from frame to frame and its distribution may be related to the speaker identity. To examine the effect of F_m, the output speech was synthesized using the estimated value of F_m and fixed values of F_m, in the range 1.0–8.0 kHz. A comparison was carried out visually using spectrograms, by objective measure of PESQ-MOS, and through informal listening. Fig. 3.10 shows spectrograms of the segment /kər/from an utterance of a female speaker. The spectrograms of the synthesized speech with estimated F_m and F_m set at 4 kHz show clear and smooth pattern of formants. Speech synthesized with F_m set at 1 kHz results in noise-like structures in place of higher formants.

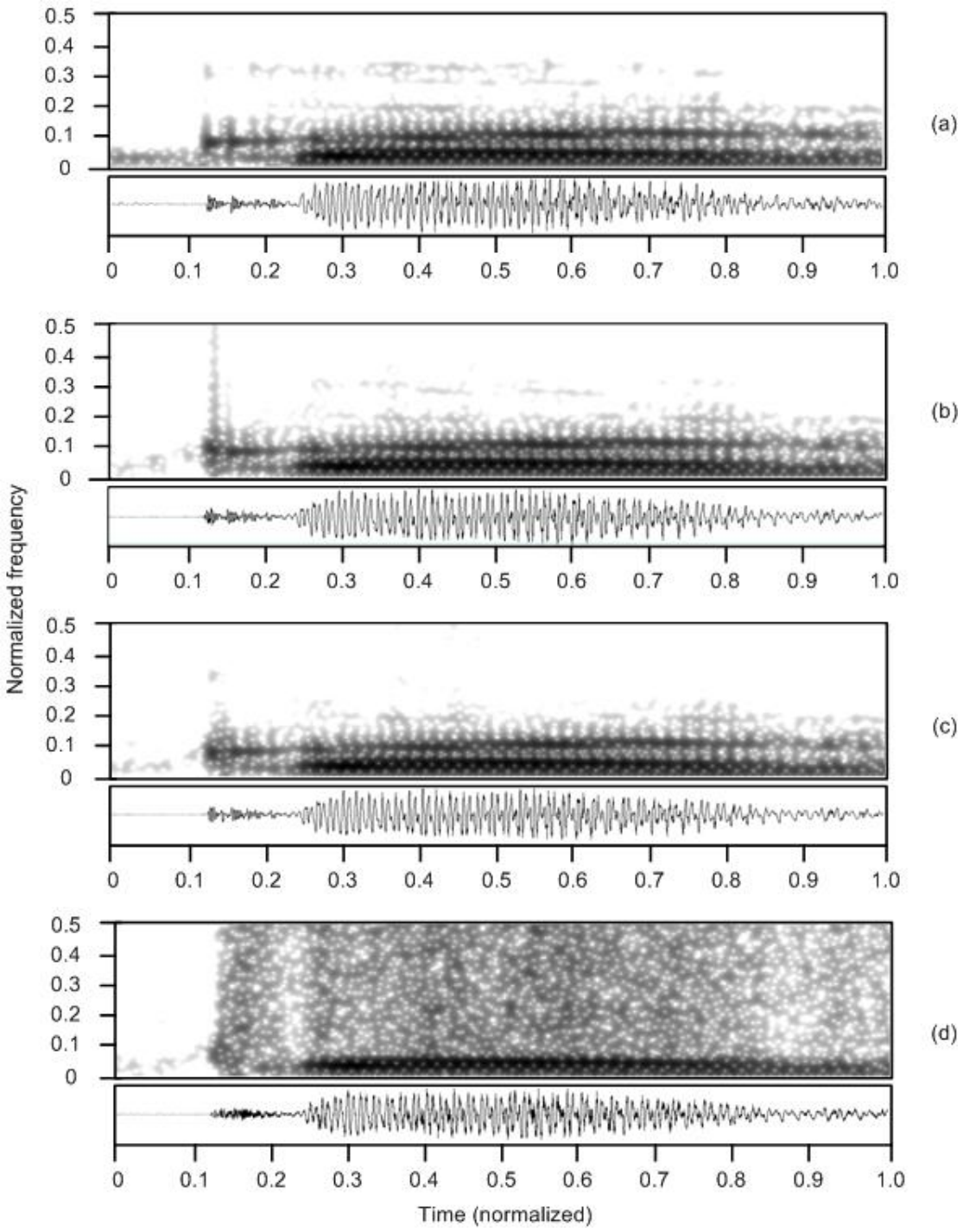

Fig. 3.10 Investigation II: Effect of F_m. Spectrograms of the segment /kə̲ɾ/(F_s = 16 kHz, duration = 0.230 s) of the Hindi utterance /d̪ʰoːbiːn̲ dʒəb soːkər ʊt̲ʰt̪i t̪oː d̪ekʰt̪i ki tʃoːkaː saːpʰ pəd̲aː həi ɔːr bərt̪ən mən̲dʒɛː hʊɛː həiːn̲/ spoken by a female speaker (F1). a) recorded, b) synth. with original F_m, c) synth. with F_m = 4 kHz, and d) synth. with F_m = 1 kHz.

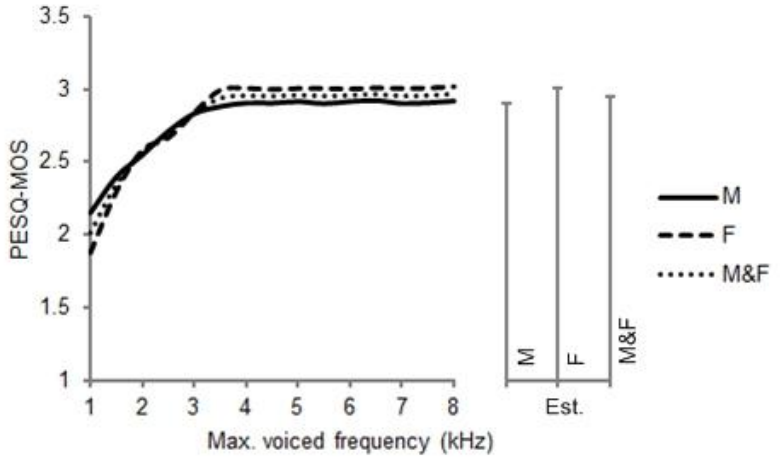

Fig. 3.11 Investigation II: PESQ-MOS test scores of the synthesized utterances (with recorded as the reference), for synthesis using different values of F_m, averaged over male (M), female (F), M&F sets of utterances for HNM-3 based analysis-synthesis.

The means and standard deviations of PESQ-MOS test scores for different values of F_m are plotted in Fig. 3.11 and also listed in Table A.2 with the last row showing the results for estimated F_m in Appendix A. With increase in F_m, the score increases, but with very small change for the values of F_m above 4 kHz. The scores for $F_m > 4$ kHz are almost the same as those with the estimated F_m. Listening the output confirmed that the speech quality became poor on decreasing F_m below 4 kHz and no appreciable change in the quality was observed for higher values of F_m .Hence it may be concluded that voice conversion framework need not involve transformation of F_m.

3.5.3 Investigation III: Effect of error in GCI estimation

Speech signal has a certain cycle-to-cycle variability in the pitch. It is quantified as jitter [219], [220], [221] and is given by

$$\text{Jitter} = \frac{\frac{1}{N-1}\sum_{n=1}^{N-1}\left|F_0(n+1)-F_0(n)\right|}{\frac{1}{N}\sum_{n=1}^{N}F_0(n)} \tag{3.39}$$

where $F_0(n)$ is the pitch frequency in frame n and N is the total number of frames. As our implementation of pitch-synchronous analysis-synthesis is based on GCIs, errors in estimation can result in increased jitter and thereby adversely affect the quality of the synthesized speech. Therefore the effect of the amount of added perturbation in estimated GCIs on quality of the synthesized speech was examined for HNM-3. Uniformly distributed random errors with zero mean were added to the estimated pitch values, resulting in

$$F_0'(n) = F_0(n) + \gamma a(n)\bar{F}_0$$

where $a(n)$ is a uniformly distributed random number in the range [-1, 1], $\bar{F}_0$ is the mean

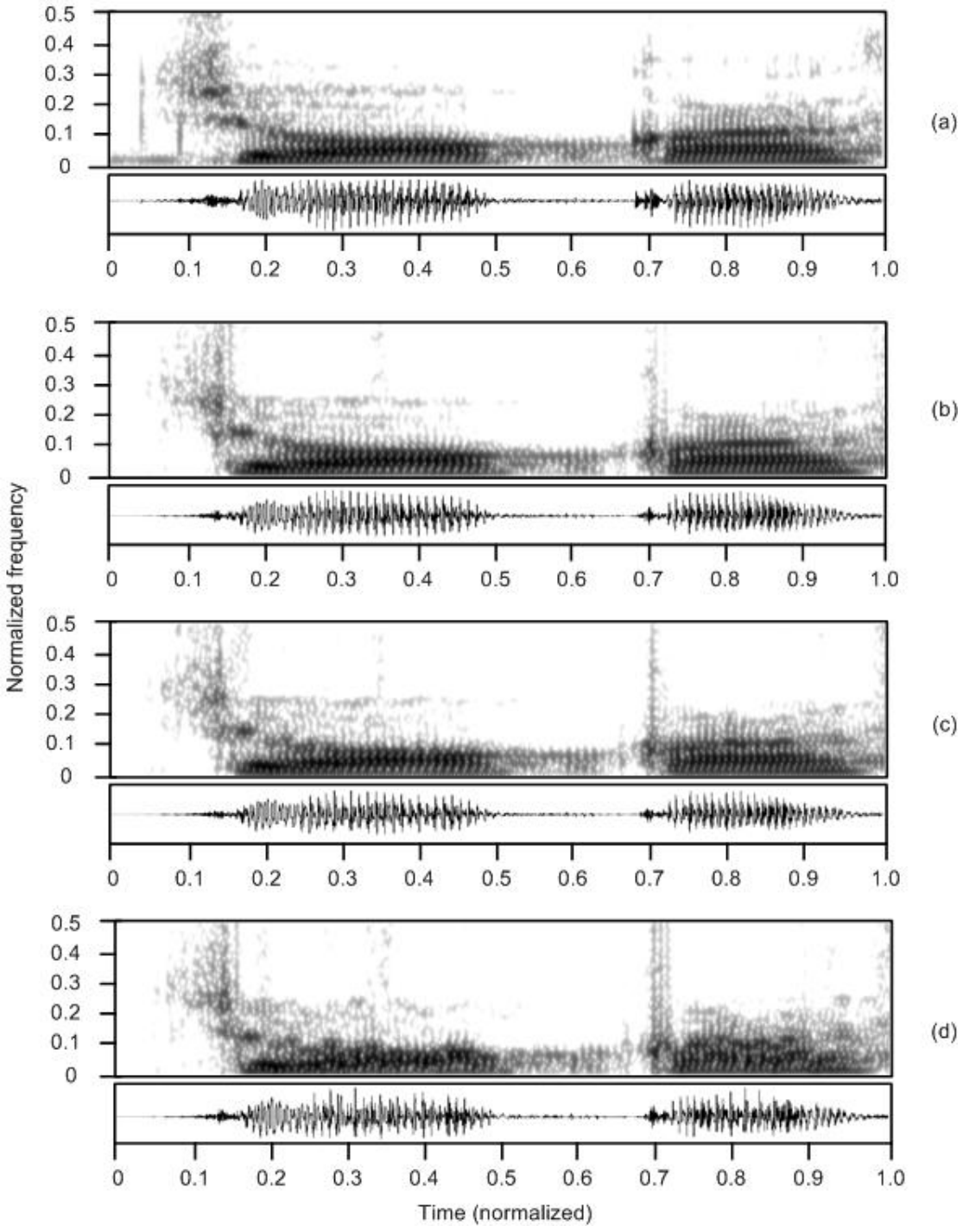

Fig. 3.12 Investigation III: Effect of jitter in GCI estimation. Spectrograms of the segment /tʃoːkaː/(F_s = 16 kHz, duration = 0.439 s)in Hindi utterance /d̪ʰoːbiːn̪ dʒəb soːkər ʊt̪ʰt̪i t̪oː d̪ekʰt̪i ki tʃoːkaː saːpʰ pəd̪aː həi ɔːr bərt̪ən mən̪dʒɛː hʊɛː həiːn̪/spoken by a male speaker (F1). a) recorded (1.58%), b) synth. with jitter = 2.92%, c) synth. with jitter = 6.10%, and d) synth. with jitter = 6.73%. The horizontal axis is normalized time and the vertical axis is normalized frequency.

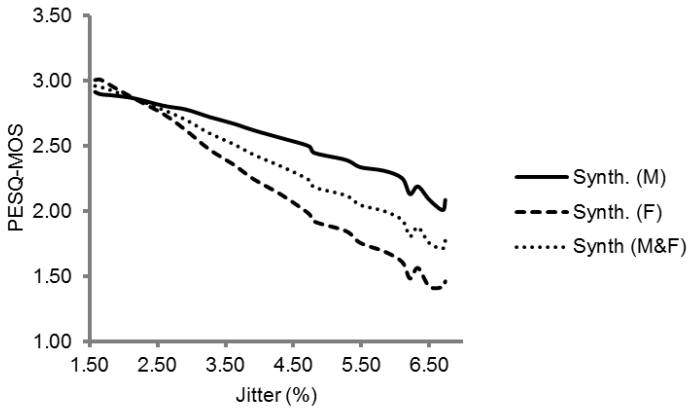

Fig. 3.13 Investigation III: PESQ-MOS test scores for the synthesized utterances (with recorded as reference) for synthesis using different amount of jitter, averaged over male (M), female (F), M&F sets of utterances for HNM-3 based analysis-synthesis, b) scatter plot of jitter in 10 vowels of two male speakers, c) error magnitude difference of GCI locations obtained from speech and EGG signals for the cardinal vowel /a/.

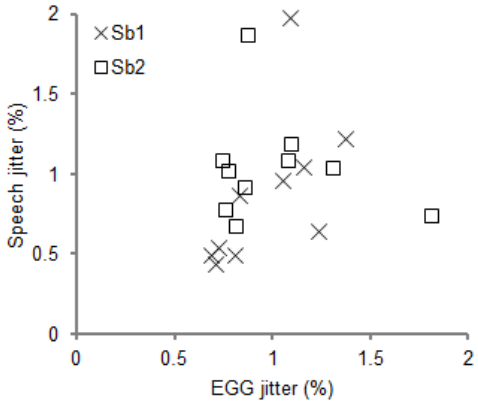

Fig. 3.14 Investigation III: Scatter plot of jitters obtained from speech and EGG signals in 10 vowels from two male speakers.

pitch, and γ is the perturbation control factor. The positions of GCIs were found in accordance with the modified pitch frequencies and HNM-3 based synthesis was carried out. A comparison was carried out using spectrograms, objective measure of PESQ-MOS, and listening.

Fig. 3.12 shows spectrograms of a segment /tʃoːkaː/ from an utterance of a female speaker (F1) for different values of jitter. Increase in jitter masks the vertical striations. Similar trends were observed in spectrograms for other utterances. The means along with the standard deviations of PESQ-MOS test scores for different values of jitter for each set of speakers are plotted in Fig. 3.13 (and also listed in Table A.3 in Appendix A). With increase in introduced jitter, the score decreases, and the effect is more visible for female speech. Degradation due to added jitter was similar to that due to additive random noise. It was more

pronounced in the utterances having longer vowel segments. The speech remained acceptable for added jitter of up to 3% (corresponding to about ±6% perturbation of the average pitch frequency). A jitter larger than 5% noticeably degraded the speech quality. These limits were approximately the same for male and female speakers.

For estimating the errors introduced by the GCI detection, speech and electro-glottograph (EGG) signals [222], [223] for 10 sustained vowels were simultaneously recorded, from two male speakers. Fig. 3.14 shows the scatter plot of the jitter in EGG signals and the corresponding speech signals. Although there is no clear relationship between the jitters as measured in the two waveforms, the ranges for the two are almost the same (< 2%). The standard deviation of the differences between the two measurements, across the utterances, is 0.36%. Thus the method used for GCI estimation may be considered as satisfactory for HNM based analysis-synthesis.

3.5.4 Investigation IV: Effect of input SNR on analysis-synthesis

Presence of noise in the input speech may introduce errors in the estimated parameters and hence it may result in distortions in the synthesized output. A comparison of the synthesized output and input was carried out using spectrograms, PESQ-MOS test, and listening.

Fig. 3.15 shows spectrograms of the input and synthesized signals of a Hindi utterance for SNR of ∝, 18 dB, 6 dB, and 4 dB. The spectrograms of the input and synthesized output are almost similar. Same pattern was observed for other speakers, for SNR down to 6 dB. The means and standard deviations of PESQ-MOS test scores noisy input and the corresponding synthesized speech (with recorded speech as reference) are plotted in Fig. 3.16 (and also listed in Table A.4 in Appendix A).The score decreases with decrease in SNR. For SNR higher than 18 dB, the noise added speech and the corresponding synthesized speech are fully intelligible and they are almost indistinguishable in quality. An interesting observation is that at lower SNR, the scores for synthesized speech are slightly higher than those of noisy input. This is possibly due to rejection of some input noise in HNM based analysis-synthesis, because the magnitude spectrum is sampled only at pitch harmonics.

3.5.5 Investigation V: Effect of phase estimation methods

In HNM, the analysis of a speech frame provides harmonic magnitude spectrum, harmonic phase spectrum, pitch frequency, GCI, voicing flag, and noise part parameters. In voice conversion, the magnitude spectrum of the source is transformed using transformation function obtained from the mapping between the source and target feature vectors aligned by DTW. Two techniques were compared for estimating the target phase. The first technique used interpolated source phase. The source phase spectrum was unwrapped along the time

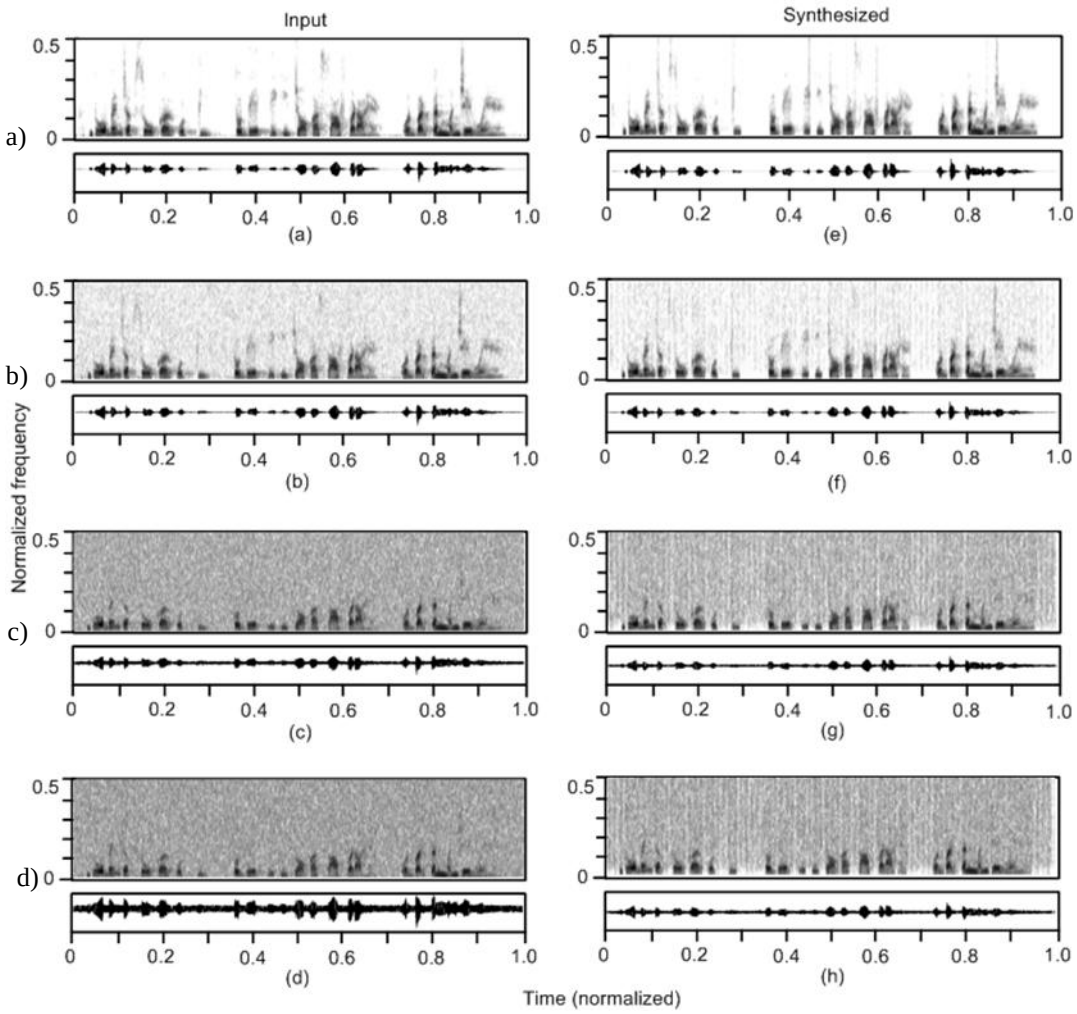

Fig. 3.15 Investigation IV: Effect of input SNR on analysis-synthesis using HNM-3. Spectrograms of the Hindi utterance (F_s = 16kHz, duration = 7.37 s) /d̪^{h}oːbiːn̪ dʒəb soːkər ʊt̪^{h}t̪i t̪oː d̪ek^{h}t̪i ki tʃoːkaː saːp^{h} pəd̪aː həi ɔːr bərt̪ən mənd̪ʒɛː huɛː həiːn̪/ spoken by a female speaker (F1) with different SNR values. a) recorded, b) 18 dB, c) 6 dB, and d) 4 dB.

and frequency axes before interpolation, as described in Subsection 3.3.2 and was interpolated at the desired target pitch harmonics.

The second technique used the phase spectrum estimated from the transformed magnitude spectrum using the method described in Subsection 3.4.2. The comparison was carried out using PESQ-MOS test and informal listening. For examining the effect of the number of MFCCs, the analysis-synthesis has also been carried out with p = 14, 18, and 22 coefficients.

The mean and standard deviations of the PESQ-MOS test scores averaged over a set of six utterances, with each set from four male and four female speakers are plotted in Fig. 3.17 (also listed in Table A.5 in Appendix A), for analysis-synthesis (AS), estimated phase (EP), and source phase (SP). The scores were slightly higher for female speakers. The scores are lower for 14 coefficients, but they are almost the same for 18 and 22 coefficients. The scores for analysis-synthesis were slightly higher than those for source and the scores for

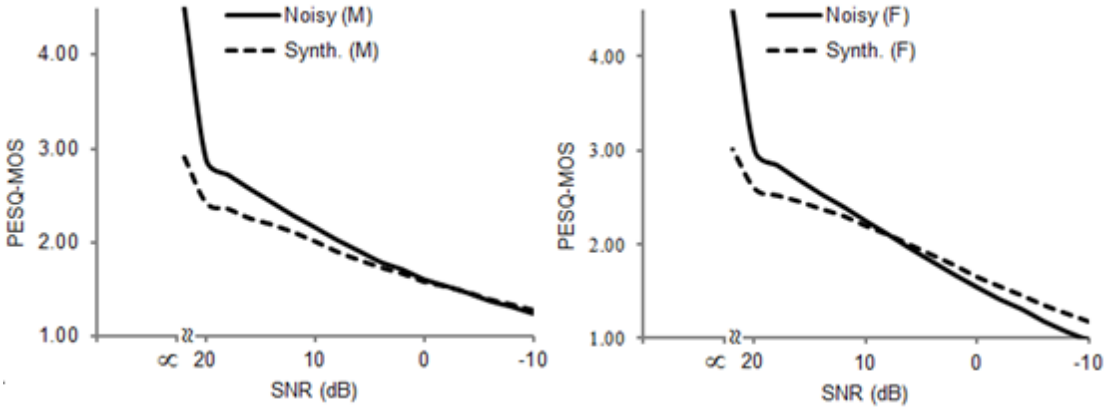

Fig. 3.16 Investigation IV: Effect of input SNR on analysis-synthesis using HNM-3. PESQ-MOS test scores (with recorded speech as reference) averaged over male (M), female (F), M&F sets of utterances for HNM-3 based analysis-synthesis.

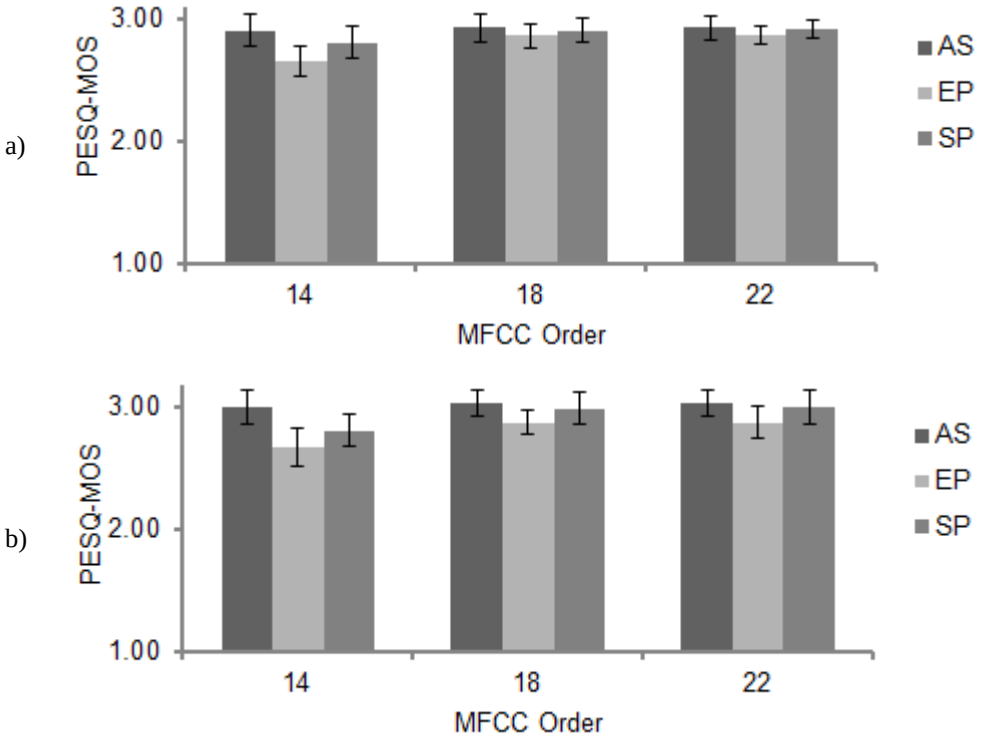

Fig. 3.17 Investigation V: Effect of estimated phase. PESQ-MOS test scores for analysis-synthesis (AS), synthesized with estimated phase (EP), and synthesized with source phase (SP), averaged across sentences and four speakers. The standard deviation is shown by error bars. a) Male speaker set, b) Female speaker set.

the estimated phase were slightly lower than those for the source phase. The differences were small and not statistically significant. The output synthesized using estimated phase was found to be perfectly intelligible but less natural in quality than the speech synthesized with analysis-synthesis and synthesized with source phase.

3.5.6 Investigation VI: Pitch and time scaling

For investigating the effect of pitch and time scaling, scaling factors were varied from 0.3 to 5. The quality of the output was examined using spectrograms and listening. Spectrograms for the signal (a) recorded, (b) pitch scaled by a factor of 0.5, (c) pitch scaled by a factor of 1.5, (d) time scaled by a factor of 0.5, and (e) time scaled by a factor of 1.5 are shown in Fig. 3.18.

The spectrograms of the pitch or time scaled speech closely resemble in formant structure to that of the unprocessed. The spectrograms (Fig. 3.18b and Fig. 3.18c) show that the pitch scaling did not alter the locations and bandwidths of the formants and their transition. In time scaled spectrograms (Fig. 3.18d and Fig. 3.18e), the formant transition is scaled as per the time scaling factor. Similar results were observed in the spectrograms of utterances from other speakers. Listening the outputs showed that the method was capable of modifying the speech by a large scaling factor For both male and female speakers, the quality and intelligibility of pitch-scaled speech remained acceptable for pitch scaling factor of 0.6–2.4. For time scaling, the speech of male speakers remained natural sounding for scaling factor of 0.5–3. For female speakers, this range was 0.6–5. For time scaling factors less than 0.5, some of the phonemes are not perceived distinctly and the intelligibility becomes poor. Satisfactory quality and intelligibility of the pitch and time scaled speech over wide range of scaling factors shows that the analysis-synthesis method used in this investigation is suitable for pitch and time scaling in cross-gender voice conversion.

3.6 Summary

The results from the investigations to examine issues related to implementation of HNM based analysis-synthesis for voice conversion can be summarized as the following.

1) The three HNM variants resulted in satisfactory quality of the synthesis output, with no significant differences. As HNM-3 involves only one set of feature vectors for the voiced frames, it is better suited as a platform for voice conversion.
2) There was no advantage of using time-varying F_m and hence its transformation during voice conversion is not needed. It may be fixed at 4 kHz.
3) An introduced jitter of more than 5% resulted in quality degradation. The effect was more pronounced in utterances with longer vowel segments. As the jitter may get introduced because of errors in detection of the GCIs, precise GCI detection is necessary for high quality synthesis of speech. The results of pitch detection were found to be satisfactory in comparison to pitch detection using an electroglottograph.
4) The analysis-synthesis method is tolerant to the presence of additive noise in the input signal.

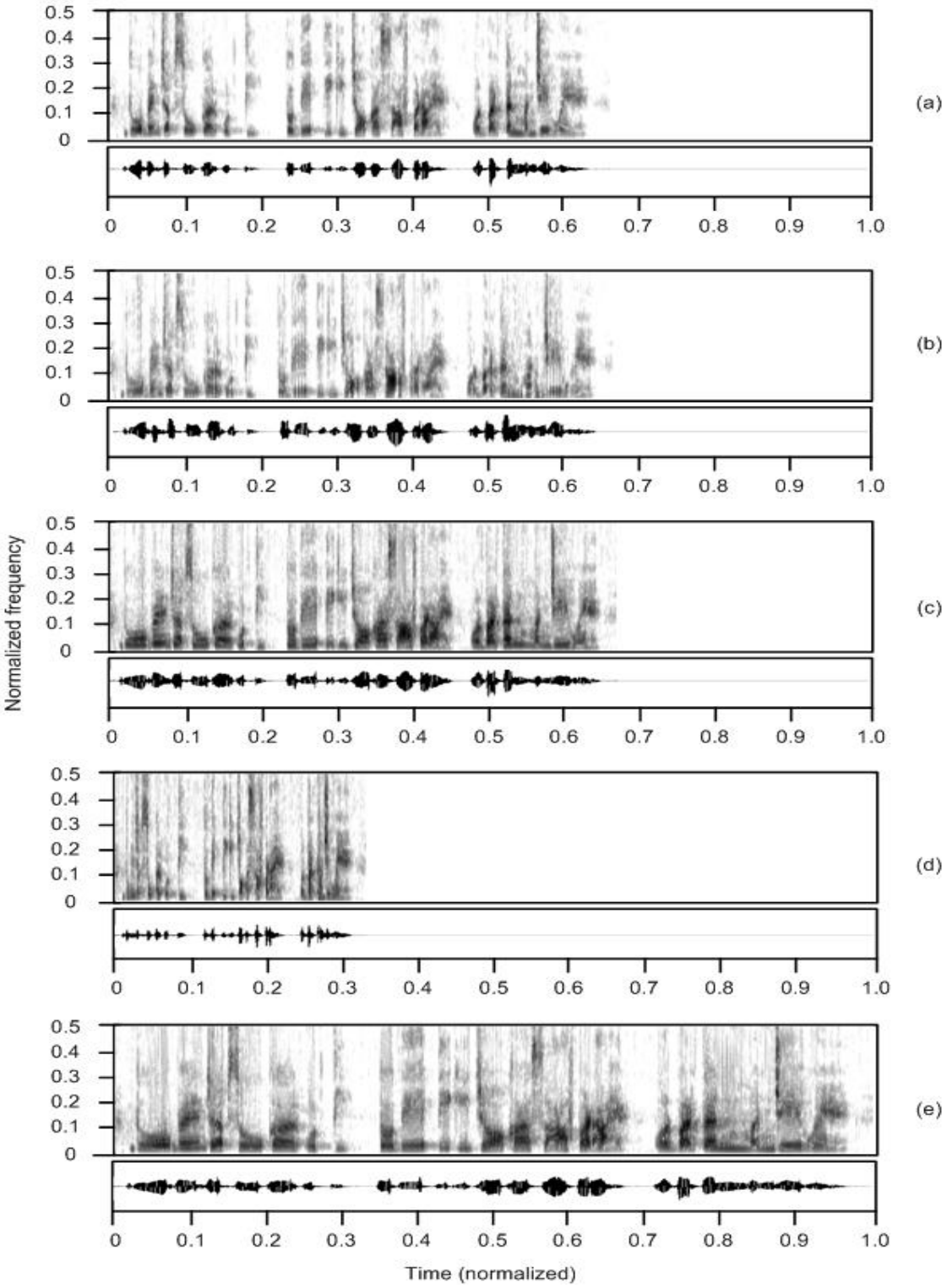

Fig. 3.18 Investigation VI: Pitch and time scaling. Spectrograms of the Hindi utterance (16 kHz, 7.37 s) /ɖ̪ʰoːbiːn dʒəb soːkər ʊ̪tʰ̪ti ʈ̪oː ɖ̪ekʰ̪ti ki tʃoːkaː saːpʰ pəɖ̲aː həi ɔːr bərʈ̪ən məɳ̲dʒɛː hʊɛː həiːn̲/ spoken by a female speaker (F1). a) recorded, b) pitch scaled by a factor of 0.5, c) pitch scaled by a factor of 1.5, d) time scaled by a factor of 0.5, and e) time scaled by a factor of 1.5.

5) The output synthesized with estimated phase was intelligible but less natural sounding than the output synthesized with source phase. As difference in the quality is small, the phase estimation may be used for voice conversion.
6) HNM-3 can be used for pitch and time scaling with a wide range of scaling factors for male and female speech. Thus it can be considered as suitable for voice conversion across speakers with different pitch ranges and speaking rates.

[blank]

Chapter 4

SPECTRAL MAPPING USING MULTIVARIATE POLYNOMIAL MODELING

4.1 Introduction

Voice conversion involves two phases: estimation of transformation function for source-to-target mapping and transformation of source speech [1], [14], [19]-[21]. In the first phase, speech signals of the source and the target speakers are converted to parametric representation in the form of feature vectors. These are used for estimating the mapping from the acoustic space of the source to that of the target. During the second phase, the given speech signal of the source is converted to feature vectors, these are modified using the transformation function, and the modified feature vectors are used for resynthesizing the speech signal. The conversion generally involves modification of the spectral as well as the prosodic parameters.

A review of different techniques used for estimating the transformation function for spectral parameters has been presented in Chapter 2. These techniques can be broadly grouped as being based on vector quantization, ANN and statistical, frequency warping, and speaker interpolation. Vector quantization uses discrete classes and hence is unable to represent the dynamic character of the speech signal. The ANN and statistical techniques need a large set of training data. The transformation also suffers from over-smoothing of the transformed spectral envelope and discontinuities across consecutive speech frames [224]. Frequency warping and speaker interpolation need different transformation function for each acoustic class. Some of these problems may be overcome by using a technique for modification of spectral characteristics for voice conversion by modeling the relationship between the acoustic spaces of the source and target speech using a single mapping.

An unknown non-linear mapping can often be approximated satisfactorily using a polynomial function of an appropriate degree. Our hypothesis is that a mapping applicable to all acoustic classes may be obtained using multivariate polynomial modeling of the relationship between the sets of parameters representing the spectral envelopes of the source and target speech signals. Such a model is expected to provide a smooth interpolation

function for transforming the source feature vectors for voice conversion using limited training data and without grouping the data into acoustic classes. In the proposed technique, each parameter for generating the target speech is modeled as a multivariate polynomial function of the parameters of the source speech. The set of these functions is obtained from the corresponding source and target feature vectors. Voice conversion of the source speech is carried out by applying the estimated mapping for modification of the spectral characteristics along with pitch and time scaling. Pitch scaling is used to match the range of the pitch in the source speech to that in the target speech. Time scaling is used to approximately match the duration of the source speech to that of the target. The modified HNM, referred to as HNM-3 in Chapter 3, is used as a platform for voice conversion as it does not need separate transformation function for noise part in voiced frames.

The details of the proposed technique and its implementation for voice conversion using parallel speech data for training are presented in this chapter. Its evaluation using objective measures and listening tests is presented in the following chapter.

4.2 Multivariate polynomial modeling

A multivariate function $g(v_1, v_2, \cdots, v_p)$ of p variables and known at N points can be approximated by a polynomial f [225]-[227]

$$g(v_{1,n}, v_{2,n}, \cdots, v_{p,n}) = f(v_{1,n}, v_{2,n}, \cdots, v_{p,n}) + \varepsilon_n, \quad 1 \le n \le N \tag{4.1}$$

where ε_n is the approximation error. Let the approximation function f be written as

$$f(v_1, v_2, \cdots, v_p) = \sum_{k=0}^{L-1} \alpha_k \phi_k(v_1, v_2, \cdots, v_p) \tag{4.2}$$

where L is the number of terms in the polynomial. For example, (4.2) for a quadratic function with 3 variables has 10 terms and can be written as

$$\begin{aligned} f(v_1, v_2, v_3) = \alpha_0 + \alpha_1 v_1 + \alpha_2 v_2 + \alpha_3 v_3 + \alpha_4 v_1^2 + \alpha_5 v_2^2 \\ + \alpha_6 v_3^2 + \alpha_7 v_1 v_2 + \alpha_8 v_2 v_3 + \alpha_9 v_3 v_1 \end{aligned} \tag{4.3}$$

By combining (4.1) and (4.2), we get a matrix equation

$$\mathbf{g} = \mathbf{A}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{4.4}$$

where vectors $\mathbf{g}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\varepsilon}$ are given by

$$\mathbf{g}^{\mathrm{T}} = \begin{bmatrix} g_1 & g_2 & g_3 & \cdots & g_N \end{bmatrix}$$

$$\boldsymbol{\alpha}^{\mathrm{T}} = \begin{bmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \cdots & \alpha_{L-1} \end{bmatrix}$$

$$\boldsymbol{\varepsilon}^{T} = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \varepsilon_3 & \cdots & \varepsilon_N \end{bmatrix}$$

Matrix $\mathbf{A}$ is $N \times L$ matrix, with the elements given as

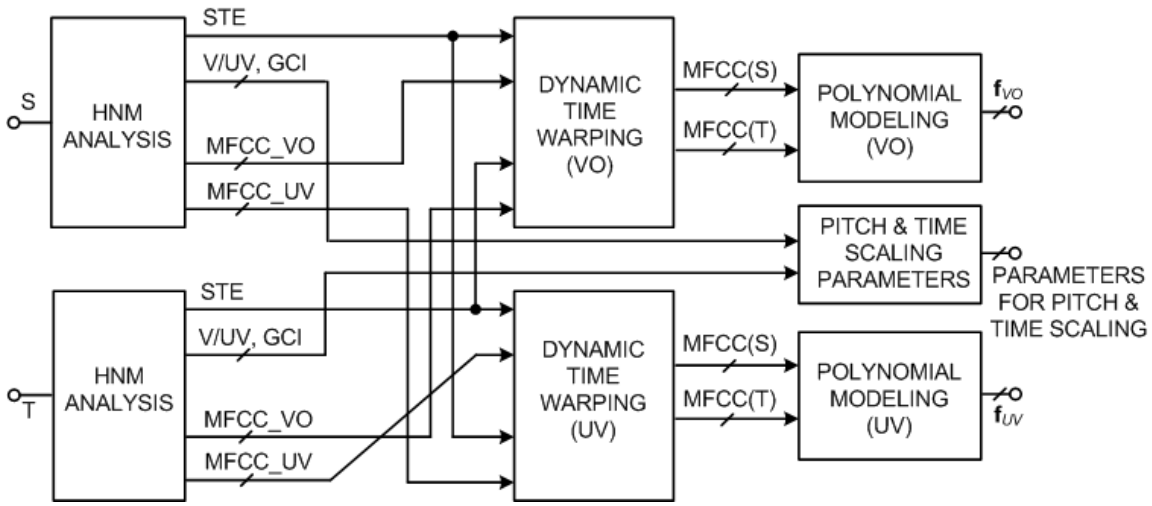

Fig. 4.1 Estimation of transformation function using HNM.

$$a_{n,k} = \phi_k(v_{1,n}, v_{2,n}, \cdots, v_{p,n}),\ 1 \le n \le N \text{ and } 0 \le k \le L-1$$

If the number of data points is greater than the number of terms in the polynomial, i.e. $N \ge L$, then coefficients α_k's can be determined for minimizing the sum of squared errors

$$E = \sum_{n=1}^{N} \left[g(v_{1,n}, v_{2,n}, \cdots, v_{p,n}) - f(v_{1,n}, v_{2,n}, \cdots, v_{p,n}) \right]^2 \tag{4.5}$$

The resulting solution can be written as

$$\boldsymbol{\alpha} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{g} \tag{4.6}$$

where $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}$ is known as pseudo–inverse of $\mathbf{A}$ [226].

Application of multivariate polynomial modeling for estimation of transformation function and transformation of the source speech is described in the following two sections.

4.3 Estimation of transformation function

The scheme for estimation of the transformation function, representing the mapping from the acoustic space of the source speaker to that of the target speaker, for voice conversion is shown in Fig. 4.1. Speech signals corresponding to the same text from the source speaker S and the target speaker T are analyzed using HNM-3, described in Section 3.4. In voiced segments, the duration of the analysis window is two pitch periods with one pitch period overlap. The analysis of the voiced frames gives GCIs and the harmonic magnitudes which are converted to discrete MFCCs. The unvoiced segments are analyzed using a 20 ms window with 50% overlap, and the magnitude spectra are converted to MFCCs. The means and standard deviations of the source and target pitch frequency distributions in log domain are calculated for pitch scaling and the ratio of the total durations of voiced segments in the source and target speech signals is calculated for time scaling.

For time alignment of the source and target frames, the frames corresponding to silence intervals, detected on the basis of short-time energy (STE), are removed. The time alignment is carried out using dynamic time warping (DTW) [228], [229]. The voiced and unvoiced frames are aligned separately. Let the source and target feature vectors be represented by $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_I)$ and $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_J)$, respectively. A grid of $I \times J$ points in a two-dimensional plane is constructed. At each grid point (i, j), the Euclidean distance $d(i, j)$ between the feature vectors $\mathbf{x}_i$ and $\mathbf{y}_j$ is calculated. A path is found for minimizing the sum of source-target distances along it. The path is searched considering three constraints. Its start and end points are (1, 1) and (I, J), respectively. For preserving the time-ordering of the feature vectors, the path has to be monotonic, i.e. at least one of the coordinates in (i, j) should increase. The step size should restrict long jumps, allowing only jumps (1, 1), (1, 0), and (0, 1). The values of i and j along the path define the time warping function between the source and target feature vectors. The first MFCC coefficient, representing the log energy, is removed before alignment to avoid any biasing due to energy fluctuations.

Due to a non-uniform representation of phonemes in the parallel speech data used for training, the time aligned feature vectors are usually in the form of clusters of different sizes which may correspond to an uneven sampling of the acoustic spaces. Decreasing the number of feature vectors based on a distance metric may help in a more even sampling of the acoustic spaces and in improving the estimation of the transformation functions. For this purpose, the feature vectors having Mahalanobis distances less than a set threshold are removed from the corresponding positions of the aligned source and target feature vectors. The threshold is empirically determined by estimating the average distance between the similar frames of vowels taken from same contexts.

We have employed polynomial modeling for deriving the mapping from the acoustic space of the source speaker to that of the target speaker represented by aligned p-dimensional feature vectors $\mathbf{x} = [x_1\, x_2\, x_3\, \ldots\, x_p]$ and $\mathbf{y} = [y_1\, y_2\, y_3 \quad y_p]$, respectively. Each element of the target feature vector is modeled as a polynomial function of all the elements of the source feature vector, i.e.

$$y_i = f_i\left(x_1, x_2, x_3, \cdots, x_p\right), \qquad 1 \le i \le p \tag{4.7}$$

Thus the mapping consists of p polynomial functions, estimated using (4.6) and aligned source and target feature vectors $(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\ldots,\mathbf{x}_N)$ and $(\mathbf{y}_1,\mathbf{y}_2,\mathbf{y}_3\ldots,\mathbf{y}_N)$ as the training data. The HNM-3 based spectral transformation for voice conversion is carried out with 20 discrete MFCCs in the voiced frames and 20 MFCCs in the unvoiced frames as the spectral parameters. The mapping obtained from the source acoustic space to the target acoustic space

is represented by two sets of transformation functions: $\mathbf{f}_{VO}$ for the voiced frames and $\mathbf{f}_{UV}$ for the unvoiced frames.

We need to select the degree of the polynomial functions for spectral transformation. A polynomial with a degree higher than that needed for representing peaks, ridges, or extremal subsurfaces in the data may introduce ripples in the approximation. Further, it increases the number of coefficients to be estimated and hence needs a larger training data. It was found that estimation of the mapping for polynomials of a degree higher than two involved ill-conditioned matrices. This phenomenon was observed even after increasing the training data. Assuming that mappings between the acoustic spaces do not have multiple extremal subsurfaces, they can be approximated using polynomials of degree one or two. Investigations for voice conversion are carried out using linear and quadratic polynomials, with three types of transformation functions: univariate linear, multivariate linear, and multivariate quadratic. Univariate linear modeling (ULM) assumes that each element of the target feature vector is a linear function of the corresponding element in the source feature vector. The target feature vector $\mathbf{y} = [y_1 y_2 y_3 \ldots y_p]$ is modeled as

$$y_i = \alpha_{0,i} + \alpha_{1,i} x_i, \; 1 \le i \le p \tag{4.8}$$

In multivariate linear modeling (MLM), each element of the target feature vector is assumed to be a linear function of all elements in the source feature vector,

$$y_i = \alpha_{0,i} + \alpha_{1,i} x_1 + \alpha_{2,i} x_2 + \cdots + \alpha_{p,i} x_p, \; 1 \le i \le p \tag{4.9}$$

In multivariate quadratic modeling (MQM), each element of the target feature vector is assumed to be a multivariate quadratic function of all elements in the source feature vector,

$$\begin{aligned} y_i = {} & \alpha_{0,i} + \alpha_{1,i} x_1 + \alpha_{2,i} x_2 + \cdots + \alpha_{p,i} x_p \\ & + \alpha_{p+1,i} x_1^2 + \alpha_{p+2,i} x_2^2 + \cdots + \alpha_{2p,i} x_p^2 \\ & + \alpha_{2p+1,i} x_1 x_2 + \alpha_{2p+2,i} x_1 x_3 + \cdots + \alpha_{L-1,i} x_{p-1} x_p, \qquad 1 \le i \le p \end{aligned} \tag{4.10}$$

where $L = 1 + 3p/2 + p^2/2$. For feature vectors with 20 components, each of the polynomial function in $\mathbf{f}_{VO}$ and $\mathbf{f}_{UV}$ has 231 coefficients.

Investigations are also carried out for studying the effect of class based mapping on the effectiveness of the transformation function. For this purpose, the feature vectors are grouped into *m* discrete classes using vector quantization [230] and a separate set of MQM functions is obtained for each acoustic class. For dividing the source feature vectors into *m* classes, *m* feature vectors are randomly selected as initial class centroids. Each source feature vector is assigned to one of the classes based on the minimum distance from the class centroids. Mean of the vectors in each class is taken as the new class centroid and class membership of the feature vectors is updated. The process is repeated until the means stop changing. Grouping of

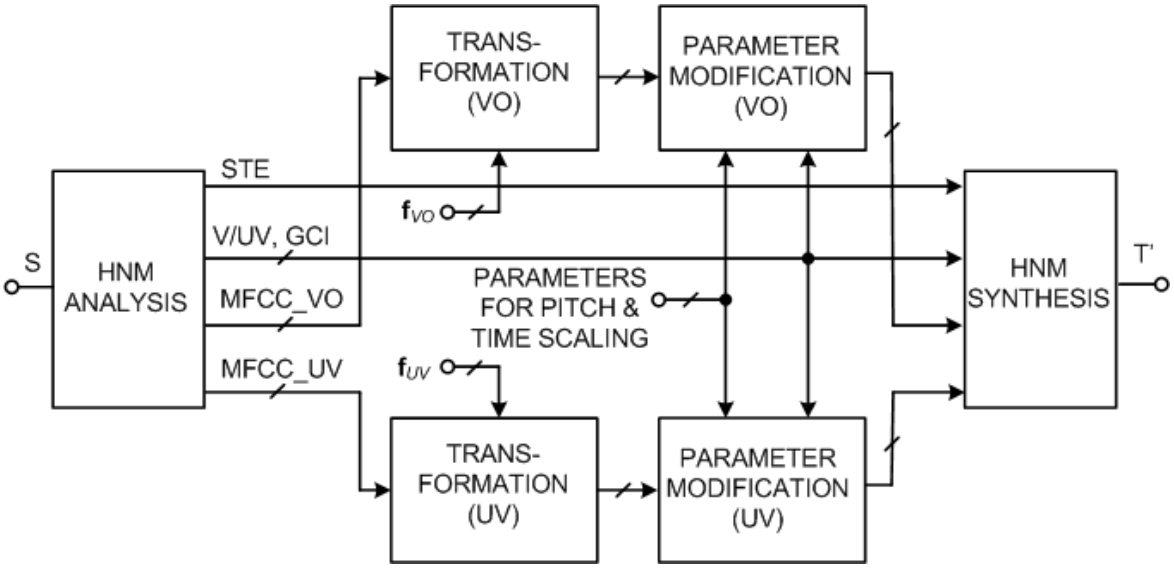

Fig. 4.2 Transformation of speech signal using HNM.

the target feature vectors follows the grouping of the corresponding source vectors because the two are aligned.

4.4 Transformation of the source speech

The scheme for transformation of the source speech to that of the target speech is shown in Fig. 4.2. The source speech is analyzed using HNM to obtain the parameters: short-time energy (STE), voicing, GCI, MFCCs for the voiced frames (MFCC_VO), and MFCCs for the unvoiced frames (MFCC_UV). The spectral parameters MFCC_VO and MFCC_UV are transformed using the transformation functions $\mathbf{f}_{VO}$ and $\mathbf{f}_{UV}$, respectively, as obtained for the given speaker pair during the training phase. The zeroth MFCC coefficient of the source feature vectors is retained in transformed MFCCs to preserve the log energy of the source speaker.

The transformed parameters are modified according to the pitch and time scaling factors as obtained from the analysis of the training data. Pitch scaling scales the pitch contour without affecting the duration. This is followed by time scaling. These two steps do not affect the spectral characteristics. Pitch scaling is carried out in accordance with the means and standard deviations of the source pitch and target pitch along log scale. The target pitch corresponding to the source pitch $P_{s,i}$ for frame i is given as

$$P_{t,i} = \exp(\mu_t + (\sigma_t / \sigma_s)(\ln(P_{s,i}) - \mu_s)) \tag{4.11}$$

where μ_s and σ_s are the mean and standard deviation of $\ln(P_{s,i})$ and μ_y and σ_y are the mean and the standard deviation of $\ln(P_{t,i})$ in the training data. The ratio of the total duration of the

voiced frames in the target speech to that in the source speech is used as the time scaling factor.

For synthesis of the target speech, the transformed MFCCs are converted to magnitude spectra as explained in Section 3.4.3. Glottal closure instants (GCIs) are marked on the synthesis axis according to the modified pitch contour. In voiced frames, the magnitude spectrum is sampled at the modified pitch harmonics. Continuous harmonic phase functions are estimated using minimum phase assumption and phase unwrapping (given in Section 3.4.2). The parameters of the voiced and unvoiced frames are interpolated at the GCI locations marked on the synthesis axis. Resynthesis from the modified parameters using HNM based synthesis provides the transformed speech. The short-time energy of the transformed speech follows that of the short-time energy of the source speech.

[blank]

Chapter 5

RESULTS AND DISCUSSION

5.1 Introduction

In the previous chapter, implementation of voice conversion system using spectral mapping based on multivariate polynomial modeling and HNM based analysis-synthesis has been presented. Results of the experiments to evaluate the voice conversion are presented and discussed in this chapter.

During the training phase, parallel set of speech data from the source and target speakers are analyzed to get aligned feature vectors. As the speech material may not be phonetically balanced, these vectors may be in the form of clusters of different sizes which may correspond to an uneven sampling of the acoustic spaces. For a more even sampling of the acoustic spaces and improving the estimation of the transformation functions, redundancy in the number of feature vectors needs to be decreased using a distance metric. The objective of the first experiment is to study the effect of distance threshold on decreasing the redundancy. The second experiment is conducted to compare the spectral mapping capabilities of the proposed MQM based transformation with the GMM based transformation, using aligned source and target feature vectors. It may be possible to estimate the source-target mapping from the feature vectors corresponding to a small number of speech segments. The third experiment is conducted to examine the effect of including and excluding different segments in the training set on the transformation functions. These experiments relate to training phase and the target speech is used as the reference for objective evaluation of the voice conversion. The fourth experiment is conducted to evaluate the quality and identity of the transformed speech using subjective listening tests. A brief description of subjective and objective evaluation methods is given in Appendix C.

5.2 Speech material

To avoid accent related bias and errors in the scores of subjective listening tests, it is desirable that the listeners and speakers belong to same group in terms of language and education. The listeners and speakers participating in our experiments were university students and they had

Hindi as their first language. There were eight speakers (4 male, 4 female, age: 20-23 years). The male speakers are referred to as M1, M2, M3, and M4 and the female speakers as F1, F2, F3, and F4. For parallel speech data, a story with 86 sentences from a children's story book in Hindi was read by each of the speakers. The material was recorded, in an acoustically treated room, using Sony ICD-PX820 audio recorder with 16 kHz sampling and 16-bit quantization. The reading duration ranged 7.9 – 11.4 min., with an average of 9.0 min. The recorded material was manually segmented into sentences having duration of 5 – 9 s. Out of them, 50 sentences were considered to be correctly articulated by all the speakers and these were selected for evaluation of voice conversion. The utterances were divided into two sets each with 25 utterances: one set for training and the other for testing. The transformation functions for polynomial modeling could be satisfactorily estimated using 10 or more utterances, while 20 or more utterances were needed for GMM. Therefore, 20 utterances from the first set were used for estimation of transformation functions for comparison of all the methods for spectral modification. All utterances of the test set were used for informal listening and six utterances were used for listening tests.

5.3 Experiment I: Redundancy of feature vectors

The transformation function is estimated from the aligned source and target feature vectors obtained from the parallel speech utterances of the source and target speakers. As the speech material used is not phonetically balanced, parameterization of the utterances in MFCC based feature vectors results in clusters with widely varying number of vectors in different clusters. As multiple instances of the feature vectors result in a highly variable weight to different vectors, estimation of transformation function can be improved by decreasing the redundancy and retaining almost similar number of vectors in each cluster. The objective of this experiment is to study the effect of the distance threshold on removal of the redundant feature vectors.

5.3.1 Material and method

For this experiment, fifty utterances for each of four male and four female speakers were taken. The utterances were converted to MFCC based feature vectors and the vectors having Euclidean distance less than a threshold were eliminated from the set. The number of vectors remaining after the reduction was calculated by varying the number of utterances from two to fifty and the distance threshold from 0 to 1.5.

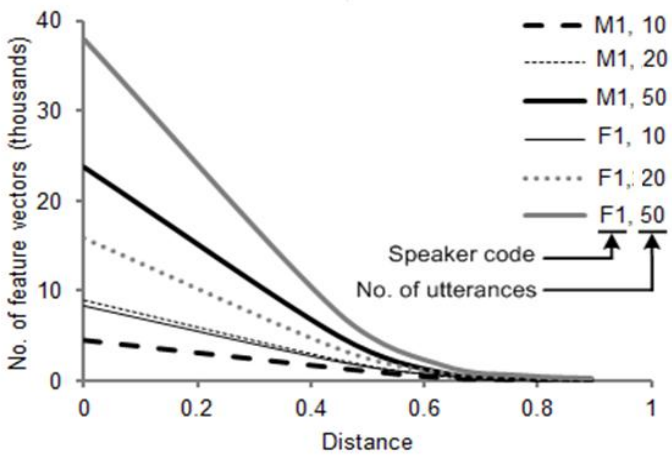

Fig. 5.1 Exp. I: Number of feature vectors (voiced and unvoiced frames) as a function of distance used in grouping.

5.3.2 Results

The number of feature vectors obtained after reduction for a male and a female speaker are shown in Fig. 5.1. The distance between the feature vectors corresponding to the same sustained vowel was observed to be generally less than 0.44, and hence feature vectors with distances smaller than this threshold may be considered to be corresponding to the same class. Thus this value may be taken as an optimal distance threshold. Removal of redundant feature vectors using it resulted in a satisfactory estimation of transformation functions for MQM using 10 or more utterances.

5.4 Experiment II: Comparison of polynomial and GMM based transformations

The accuracy of the mapping derived from the given set of feature vectors depends upon two factors: samples of the source and target acoustic spaces available for training in the form of feature vectors and the method used for the approximation of the mapping between the acoustic spaces. A total of six methods are compared in this experiment for estimating the transformation function. Five of these methods are polynomial modeling based: univariate linear modeling (ULM), multivariate linear modeling (MLM), multivariate quadratic modeling (MQM), 32 classes based multivariate quadratic modeling (MQM32), and 64 classes based multivariate quadratic modeling (MQM64). The sixth method is GMM based with 64 mixture components and diagonal covariance matrix (as described in Appendix B).

5.4.1 Material and method

The testing material consisted of two sets, each with 25 utterances from four pairs of speakers (M1-M2, F1-F2, M3-F3, and F4-M4). The utterances from the first set were used for training and six utterances from the second set were used for testing. The utterances were analyzed

Table 5.1. Exp. II: Cepstral Mahalanobis distance between the source and target (ST) and the target and transformed (TT') for different speaker pairs (mean and standard deviation for six utterances).

Distance	M1-M2		F1-F2		M3-F3		F4-M4		Avg.
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	
ST	5.61	0.96	5.72	1.04	6.12	1.06	5.86	1.12	5.83
TT'-ULM	4.07	0.77	4.30	0.81	4.48	0.85	4.03	0.79	4.22
TT'-MLM	3.89	0.78	4.05	0.83	4.22	0.86	3.75	0.77	3.98
TT'-MQM	3.74	0.77	3.83	0.83	3.99	0.86	3.61	0.76	3.79
TT'-MQM32	1.62	0.91	1.68	0.93	1.87	0.97	1.47	0.74	1.66
TT'-MQM64	0.24	0.69	0.74	1.06	0.57	1.01	0.27	0.71	0.46
TT'-GMM	3.47	0.89	3.54	0.92	3.61	1.01	3.40	0.85	3.51

using HNM, and MFCC-based feature vectors were obtained. The feature vectors were aligned using DTW. Redundancy reduction was applied in two steps. First all the source vectors with intervening Euclidean distance below a threshold were removed. Subsequently the same process was applied on the target vectors. The remaining aligned source-target pairs of feature vectors were used for training. Each parameter for generating the target speech is modeled as a multivariate polynomial function of all the parameters of the source speech, and the set of these polynomial functions is obtained by analyzing a set of time aligned source and target frames. ULM, MLM, MQM, and GMM based transformation functions were estimated. MQM32 and MQM64 based transformation functions were estimated by dividing the total feature vectors into 32 and 64 classes using k-means algorithm. The source utterances were transformed using these transformation functions and evaluation was carried out using cepstral Mahalanobis distance between the source and target (ST) and the target and transformed source (TT'). The distances were averaged across the utterances in testing set of utterances. Actual CPU time required for the estimation of each transformation function on a PC was also observed. The transformation functions for polynomial modeling could be satisfactorily estimated using 10 or more utterances, while 20 or more utterances were needed for GMM. Therefore, 20 utterances were used for estimation of transformation functions for comparison of all the methods for spectral modification.

5.4.2 Results

Mean and standard deviation of the distances between the target (T) and source (S) speech and the target and transformed speech (T') are given in Table 5.1 and plotted in Fig. 5.2. All the mapping methods resulted in a decrease in the distance for all the speaker pairs. The decreases in the distance, averaged across the speaker pairs, were 28%, 32%, 35%, 72%, 92%, and 40%, for the spectral modifications using ULM, MLM, MQM, MQM32, MQM64, and GMM, respectively. Thus on the basis of decrease in the distance, the methods can be ranked as MQM64 > MQM32 > GMM > MQM > MLM > ULM.

Informal listening showed that the output of ULM based transformation was not fully intelligible and did not sound natural. The outputs of MQM32 and MQM64 had distinct audible disturbances affecting the intelligibility, indicating that dividing feature vectors into classes adversely affected the quality of the output speech. The outputs of MLM, GMM, and MQM based transformation were intelligible and sounded natural; and MQM appeared to be relatively better in transforming the identity as compared to the other methods.

Computation time (actual CPU time) on a PC (Intel 2.53-GHz 64-bit i3-380M processor, 4 GB RAM, 320 GB HDD) for estimation of the transformation function from the training data was 18.2 s for MQM. With reference to this time, relative times for ULM, MLM, MQM32, MQM64, and GMM were 0.945, 0.947, 1.390, 1.620, and 82.51, respectively. Thus although the decrease in the spectral distance for MQM and GMM is almost similar, the computation time required for estimation of GMM based transformation was about 83 times that for MQM. ULM and MLM resulted in only a slight computational advantage over MQM, but resulted in poor speech quality. Differences in computation time for transformation were not significant.

5.5 Experiment III: Interpolation capabilities of MQM

MQM based voice conversion assumes each component in the feature vector of the target speaker to be a multivariate function of every component of the feature vector of the source speaker. Once the transformation function has been obtained from the aligned source and the target feature vectors, the function is used to transform source feature vectors to corresponding vectors in the acoustic space of the target speaker. The process assumes the feature vectors used for the estimation of the transformation functions to be uniformly distributed in the acoustic spaces. The objective of this experiment is to examine the interpolation capabilities of the MQM by estimating the transformation function using a limited number of speech segments, particularly on the transformation of speech segments not included in the training data.

5.5.1 Material and method

The testing material consisted of two sets of 54 phonemic segments from the utterances of four speaker pairs (M1-M2, F1-F2, M3-F3, and F4-M4). The segments were obtained by manually marking them in the utterances. One set was used for training and the other set was used for testing. The list of the speech segments is shown in Table 5.2. These segments were considered to be representative of the different segments appearing in the utterances. Some of the segments have same labels but they occur in different contexts. The investigations were carried out by estimating 55 sets of transformation functions after redundancy reduction (as described in Subsection 5.4.1). The set Fn0 was estimated using all the speech segments from

the training set, while other functions were estimated by deleting one of the segments (*m* for Fn*m*). The estimation process consisted of HNM analysis, DTW, and MQM coefficients calculation. During testing, the speech segments from the testing set were analyzed using HNM and MFCC based feature vectors were obtained. These vectors were transformed using the respective transformation functions. Evaluation was carried out using cepstral Mahalanobis distances between the source and target (ST) and the target and transformed source (TT'). The distances were normalized by the corresponding source-target distances and averaged across the four speaker pairs for each speech segment.

Table 5.2. Exp. III: Speech segments used for training and testing.

Sr. No.	Segment label	Description	Sr. No.	Segment label	Description
1	aa1	/a:/$_1$ long	28	ih	/i/$_2$
2	aa2	/a:/$_2$ long	29	ix1	/i/$_1$ centralized
3	ae	/ʌ/$_1$ long	30	ix2	/i/$_2$ centralized
4	ah	/ʌ/$_1$ short	31	iy1	/i/$_1$
5	ao	/ɔ/$_1$	32	iy2	/i/$_2$
6	ao	/ɔ/$_2$	33	jh	/dʒ/$_1$
7	ax	/ʌ/$_2$ short	34	k	/k/
8	axr	/ʌ/$_3$ short	35	kcl	/k/ with no release
9	ay	/a:i/	36	l	/l/$_2$
10	b	/b/	37	m	/m/
11	bcl	/b/ with no release	38	n	/n/
12	ch	/tʃ/	39	ow	/o/
13	d	/ɖ/ release	40	p	/p/
14	dcl	/ɖ/ closure	41	pcl	/p/ closure
15	dh	/d/	42	q	/q/ glottal stop
16	dx	/t/ flap or tap	43	r	/r/
17	eh	/e/	44	s	/s/
18	el	/l/$_1$	45	sh	/ʃ/
19	epi	epenthetic silence	46	t	/t/ release
20	ey	/ei/$_1$	47	tcl	/t/ with no release
21	ey	/ei/$_2$	48	uh	/u/$_1$
22	f	/p^h/	49	uw	/u/$_2$
23	g	/g/	50	ux	/u/$_3$
24	gcl	/g/ with no release	51	v	/v/$_1$
25	hh	/h/ glottal u/v fricative	52	w	/v/$_2$
26	hv	/h/ glottal voiced	53	j	/j/
27	ih	/i/$_1$	54	z	/dʒ/$_2$

a)

b)

c)

c)

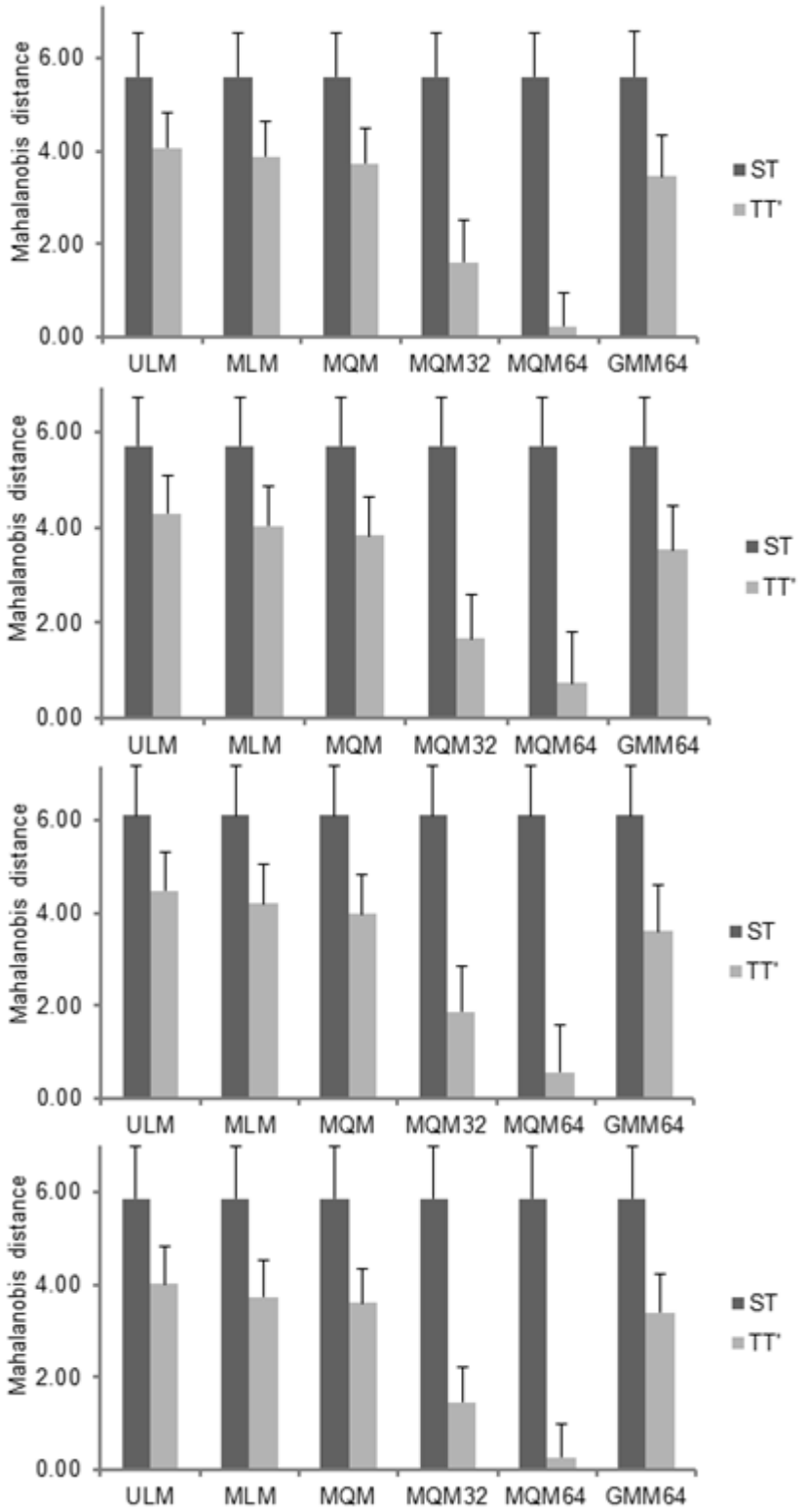

Fig. 5.2 Exp. II: Cepstral Mahalanobis distance between the source/target (ST) and target-transformed/target (TT′) pairs. a) M1-M2, b) F1-F2, c) M3-F3, d) F4-M4.

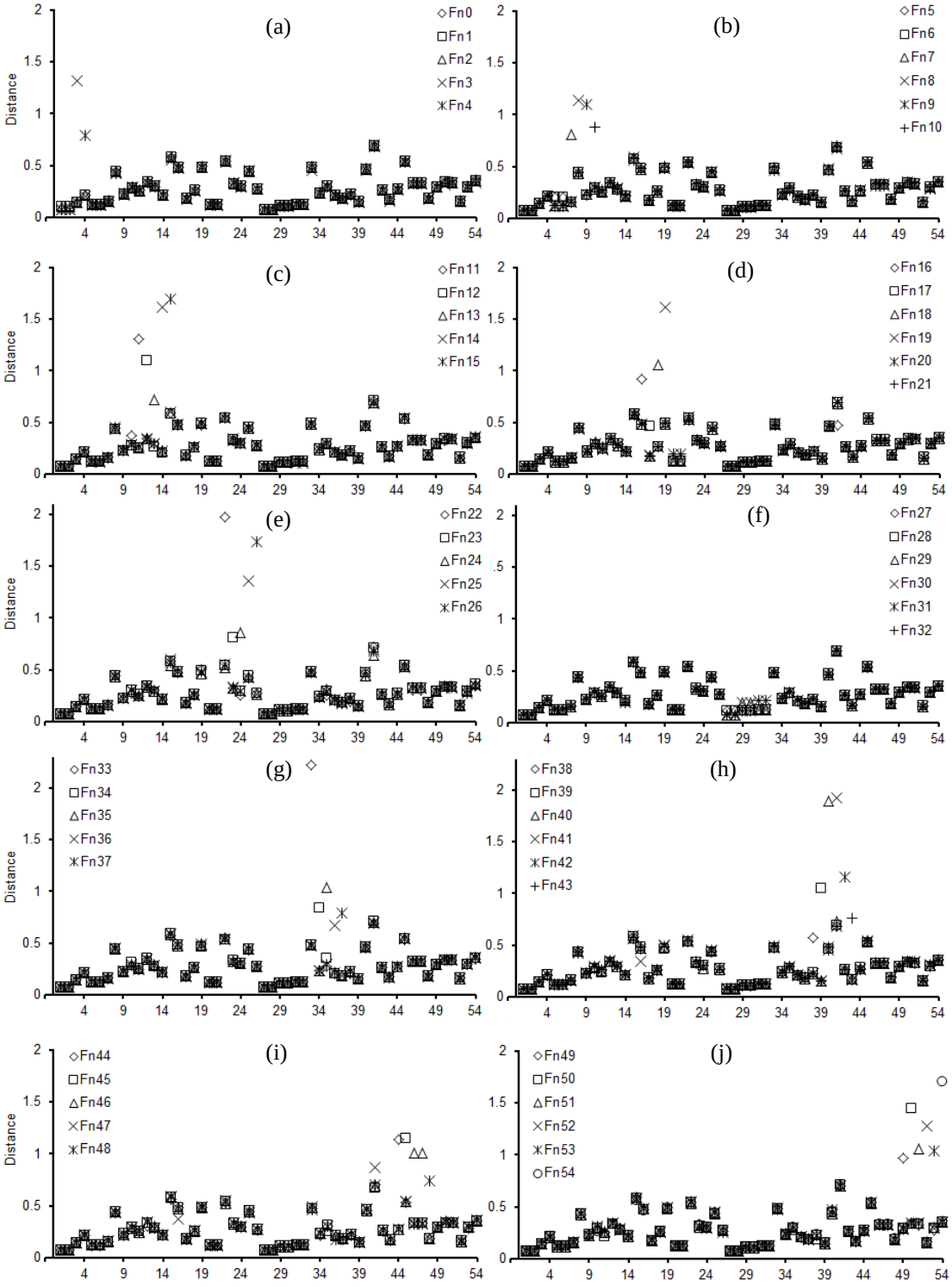

Fig. 5.3 Exp. III: Normalized average cepstral Mahalanobis distance between the target and transformed source for 54 speech segments using sets of transformation functions Fn0 to Fn54. The x-axis denotes the speech segment number.

5.5.2 Results

The average normalized cepstral Mahalanobis distances for each speech segments obtained from each of the 55 sets of transformation functions are shown in Fig. 5.3. The speech segments are along x-axis and the distances along y-axis. For transformation using Fn0, which was obtained with all the segments in the training data, the distances for all the segments are small and below 0.6. For transformation using other sets of functions (Fn1–Fn54), the distances increase for the speech segments which were not present during the training. For example, in Fig. 5.3(a), the distances are very large for the speech segments 3 and 4. The increase in the distances for fricative, affricate, and plosives are much more than those for vowels. These results show that if feature vectors corresponding to such speech segments are not available in the training data, transformation function will not be valid for them. The experiment also shows that MQM based transformation function can be estimated using a limited training data if it contains speech segments similar to all speech segments likely to occur during the testing phase.

5.6 Experiment IV: Subjective evaluation of MQM using MOS and XAB

Actual source-to-target transformation for voice conversion involves spectral modification along with the pitch and time scaling. Listening tests were conducted for subjective evaluation using MOS test for quality [231] and XAB test for identity [51] of the transformed speech. The evaluation was carried out only for MQM based spectral modification, because the results of the objective evaluation showed it to be better than the other methods. The tests were conducted to examine the contribution of spectral modification and pitch scaling separately and together.

5.6.1 Material and method

Two sets each of 25 utterances of four male and four female speakers were available for use as training and test material for this experiment. From the first set, 20 utterances were used as the parallel speech training data. Six utterances from the second set were used for listening tests. The utterances for training were analyzed using HNM, and a set of MQM-based transformation functions was obtained, after redundancy reduction (as described in Subsection 5.4.1), for each pair of speakers.

Three types of modifications were applied on the source utterances. The first modification involved pitch scaling. In the second modification, only spectral modification was carried out. In the third modification, both spectral modification and pitch scaling were applied. Time scaling was used in all three modifications, with the ratio of the total duration of voiced frames in the target speech to that in the source speech in the training data as the

time scaling factor. These modifications are referred to as pitch scaling (PS), spectral modification (SM), and spectral modification with pitch scaling (SMPS). Speech signals used in the listening tests are available online [232].

The XAB and MOS tests were conducted using a PC-based automated experimental setup with a graphical user interface (GUI) for presentation of the stimuli and recording of the responses. Each presentation involved three stimuli: X, A, B. Utterance corresponding to S, T, or T' was randomly selected as X. Either S or T was randomly selected as A and the other one was selected as B. The subject listened to the three stimuli (more than once if needed for finalizing the response) and responded by pressing the corresponding buttons on the GUI. For XAB, the subject responded by selecting either A or B as the best match to X. Percentage score for labeling the presented stimuli as the target was calculated as the XAB score. For MOS, the subjects assigned score to the quality of the stimulus X on 1–5 scale (1: bad, 2: poor, 3: fair, 4: good, 5: excellent). Instructions to the subjects participating in the listening tests are given in Appendix D.

Six normal-hearing subjects (4 male, 2 female, age: 24-27 years), participated as subjects in the listening tests. All were university students with Hindi as their first language and thus belonged to the same group as the speakers in terms of language and education. In a test, each stimulus appeared four times and the test was conducted for four speaker pairs (M1-M2, F1-F2, M3-F3, and F4-M4). For each listener, there were 480 presentations: 4 speaker pairs × 6 utterances × 4 repetitions × 5 types of stimuli (S, T, T'-PS, T'-SM, T'-SMPS). The scores were averaged across 24 presentations (6 utterances × 4 repetitions). To reduce the bias due to practice or fatigue, presentation order for different types of test utterances was randomized across the subjects.

5.6.2 Results

Listening to the transformed utterances showed that modification by applying pitch scaling and spectral modification individually resulted in speech identity being converted from source towards the target, but the modified utterances could be perceived as different from the target. Combination of spectral modification and pitch scaling made the identity of the modified utterances almost the same as that of the target. There were small audible distortions, particularly in cross-gender conversions, involving large pitch scaling factors. Some of these distortions may have also resulted due to errors in GCI detection, phase discontinuities, and artifacts related to time scaling.

The results of MOS test for the six subjects along with the means and standard deviations are given in Table 5.3. The means across the subjects are plotted in Fig. 5.4. Mean scores for source and target utterances are 4.94 and higher. Mean scores for PS utterances are

Table 5.3. Exp. IV: MOS test scores for the sets of source (S), target (T), pitch scaled (PS), spectral modification (SM), and spectral modification and pitch scaled (SMPS) utterances.

Speaker Pair	Speech	MOS score							
		Sb1	Sb2	Sb3	Sb4	Sb5	Sb6	Mean	SD
M1–M2	S	4.88	4.96	4.92	4.96	4.92	5.00	4.94	0.04
	T	5.00	4.92	4.92	5.00	4.96	5.00	4.97	0.04
	PS	3.08	2.96	3.08	3.04	3.00	3.00	3.03	0.05
	SM	2.42	2.46	2.50	2.54	2.42	2.46	2.47	0.04
	SMPS	2.63	2.71	2.63	2.58	2.83	2.96	2.72	0.13
F1–F2	S	4.92	5.00	4.96	4.92	5.00	5.00	4.97	0.04
	T	4.96	5.00	5.00	5.00	4.96	5.00	4.99	0.02
	PS	3.04	3.08	2.92	3.08	3.00	2.92	3.01	0.07
	SM	2.08	2.13	2.17	2.42	2.08	2.00	2.15	0.13
	SMPS	2.54	2.54	2.83	2.67	2.58	2.50	2.61	0.11
M3–F3	S	5.00	4.92	5.00	4.96	5.00	5.00	4.98	0.03
	T	5.00	4.96	5.00	5.00	4.96	5.00	4.99	0.02
	PS	3.00	3.00	3.04	3.13	2.92	3.08	3.03	0.07
	SM	2.08	2.00	2.13	2.04	2.00	2.08	2.06	0.05
	SMPS	3.17	2.92	3.08	2.92	3.00	3.00	3.01	0.09
F4–M4	S	4.92	4.96	5.00	5.00	5.00	4.96	4.97	0.03
	T	4.96	4.92	5.00	5.00	5.00	4.96	4.97	0.03
	PS	2.96	3.13	3.00	3.08	2.96	3.00	3.02	0.06
	SM	2.00	2.17	1.96	1.83	2.04	2.21	2.03	0.13
	SMPS	3.25	2.42	3.13	3.17	3.13	3.13	3.03	0.28

3.03, 3.01, 3.03, and 3.02 for M1-M2, F1-F2, M3-F3, and F4-M4 conversions. The corresponding mean scores for SM utterances were 2.47, 2.15, 2.06, and 2.03. These results indicate the suitability of HNM platform as a suitable platform for pitch scaling and spectral modification. They also show that spectral modification without appropriate pitch scaling degrades the perceived quality. The mean scores for M1-M2, F1-F2, M3-F3, and F4-M4 conversions using SMPS were 2.72, 2.61, 3.01, and 3.03. The MOS test scores for SMPS using MQM are similar to or slightly better than results reported earlier for GMM based voice conversion [20], [94], [156].

Subject-wise XAB test scores along with means and standard deviations for labeling of the test utterances (source, target, transformed speech) as the target utterances are given in Table 5.4 for different speaker pairs. The mean scores are plotted in Fig. 5.5. For unmodified speech, there were some errors (up to 17%) in speaker identification in the same-gender tests, but there were no such errors in cross-gender tests. The mean scores for correctly identifying the target were 90%, 87%, 100%, and 100% for speaker pairs M1-M2, F1-F2, M3-F3, and

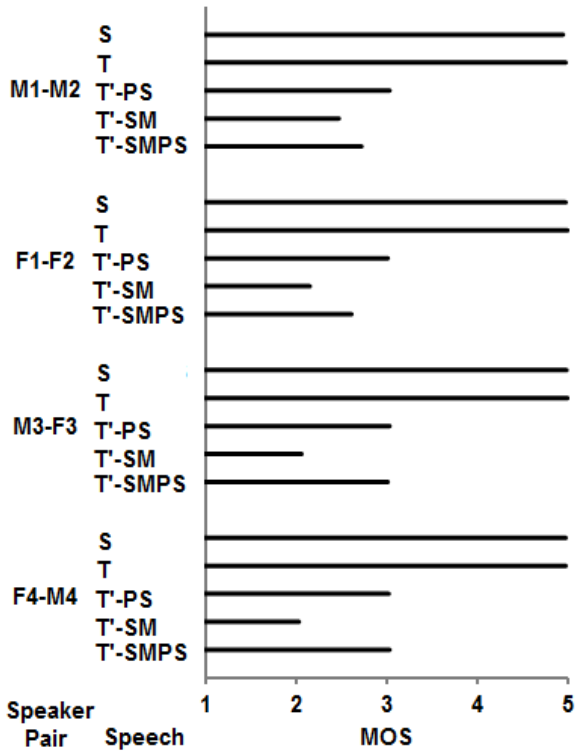

Fig. 5.4 Exp. IV: MOS test scores for the sets of source (S), target (T), pitch scaled (PS), spectral modification (SM), and spectral modification and pitch scaled (SMPS) utterances.

F4-M4, respectively. Inter-subject variability was very low. Scores for identifying the modified speech as the target showed that identification of the pitch scaled utterances and the spectrally modified utterances as the target speech was less than 14% and 6%, respectively, indicating that either of these modifications by itself is not sufficient for the converted speech being identified as the target. Identification of transformed utterances obtained by spectral modification along with pitch scaling as the target speech was 93% for the same-gender conversion and 100% for cross-gender conversion. Thus the results show that the combination of spectral modification and pitch scaling resulted in successful voice conversion.

5.7 Discussion

Estimations of transformation functions for multivariate polynomial models with a degree higher than two were found to involve ill-conditioned matrices. Hence investigations were carried out for univariate linear model (ULM), multivariate linear model (MLM), and multivariate quadratic model (MQM). Multivariate quadratic modeling was also investigated by applying the technique by dividing the feature vectors into classes using k-means algorithm, with separate transformation function estimated for each class. During conversion, the feature vector was first assigned to a class and then converted by using the corresponding transformation function. Transformations using 32 and 64 classes are referred to as MQM32

Table 5.4. Exp. IV: XAB scores for the sets of source (S), target (T), pitch scaled (PS), spectral modification (SM), and spectral modification and pitch scaled (SMPS) utterances.

Speaker pair	Speech	XAB scores for target labeling (%)							
		Sb1	Sb2	Sb3	Sb4	Sb5	Sb6	Mean	SD
M1–M2	S	13	0	0	8	4	17	7	6
	T	96	83	96	88	83	96	90	6
	PS	13	17	17	4	13	8	12	4
	SM	4	8	0	4	8	4	5	3
	SMPS	96	88	92	96	100	92	94	4
F1–F2	S	8	4	0	0	4	13	5	4
	T	83	88	83	92	83	96	87	5
	PS	17	8	4	17	13	4	10	5
	SM	8	13	8	8	0	13	8	4
	SMPS	92	92	92	92	92	92	92	0
M3–F3	S	0	0	0	0	0	0	0	0
	T	100	100	100	100	100	100	100	0
	PS	17	8	8	13	13	17	13	3
	SM	4	8	8	4	8	4	6	2
	SMPS	100	100	100	100	100	100	100	0
F4–M4	S	0	0	0	0	0	0	0	0
	T	100	100	100	100	100	100	100	0
	PS	13	17	17	8	13	17	14	3
	SM	8	8	4	4	8	4	6	2
	SMPS	100	100	100	100	100	100	100	0

and MQM64. Evaluation was carried out for four speaker pairs (male-male, female-female, male-female, and female-male), using a set of four experiments.

The first experiment showed that removal of redundant feature vectors, using a distance threshold, could be used for achieving a more uniform sampling of acoustic spaces and improving the estimation of transformation functions. Transformation functions were found to stabilize for training data consisting of ten or more sentences.

In the second experiment, transformation methods were compared by listening to the transformed speech, computation time for estimation of the transformation function, and the objective measure of the cepstral Mahalanobis distance of the transformed speech from the target speech with reference to the source-target distance. ULM and MLM involved slightly less computation time as compared to MQM. Listening of converted speech showed that ULM resulted in poor quality speech, while the outputs from MLM and MQM were natural sounding. MQM resulted in the highest decrease in the distance for all speaker pairs. MQM64 and MQM32 took longer computation time as compared to MQM, and they resulted in larger decrease in the distance between the target and the transformed speech. But MQM32 and MQM64 outputs had distinct audible distortions affecting the intelligibility and naturalness, indicating that dividing feature vectors into classes adversely affected the transformation.

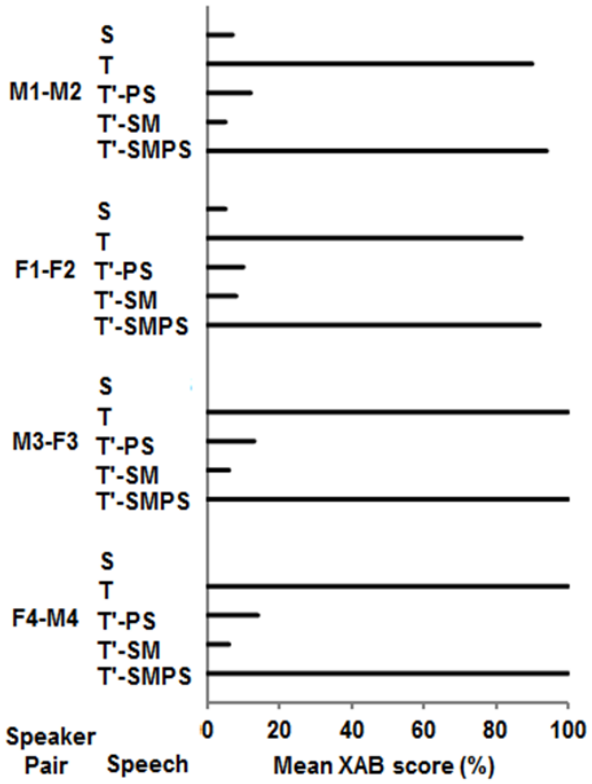

Fig. 5.5 Exp. IV: XAB scores for the sets of source (S), target (T), pitch scaled (PS), spectral modification (SM), and spectral modification and pitch scaled (SMPS) utterances.

A weighted sum of the spectral parameters obtained from transformations for different classes and use of other clustering methods may improve the quality. However, it increases the number of coefficients to be estimated and therefore will require a larger training data. Thus the results show that out of all the investigated models, MQM is the most suited one for spectral modification. A comparison was also made with GMM based transformation (64 mixture components) using the same training and test data. It was found that output quality and decrease in target-transformed distance with GMM and MQM were almost similar, but MQM needed a much shorter computation time (less than 1/83) for estimation of the transformation function.

The results of the third experiment showed that the parallel training data should have adequate representation of feature vectors corresponding to different acoustic classes. There were relatively large errors during the transformation of the speech segments not represented in the training data. The results also indicate that MQM based transformation function can be estimated by using a limited training data if it contains speech segments similar to all speech segments likely to occur during the testing phase.

In the fourth experiment, the quality and identity of the MQM based transformed speech was assessed using MOS and XAB tests for same-gender and cross-gender voice conversion, with the mapping obtained from parallel speech data consisting of twenty sentences. The MOS test scores for SMPS were higher than those for SM, but lower than those for PS. The quality degradation during voice conversion may be attributed to the discontinuities in transformed magnitude spectra and in the phases obtained from them. Pitch scaling does not involve spectral modification and hence the scores are consistently better than those for spectral modification. The quality of spectrally modified speech improves if the pitch is also modified, because it restores a more natural relationship between the pitch and vocal tract parameters. It is expected that use of methods for reducing the effect of phase discontinuities during pitch-synchronous analysis-synthesis will help in improving the quality of speech after voice conversion. Some of the audible distortions may have also resulted due to uniform time scaling during voiced segments. The XAB scores showed that combination of spectral modification and pitch scaling resulted in transformed speech having almost the same identity as the target. Averaged across listeners, the scores for identification of transformed speech as the target were 93% for same-gender conversion and 100% for cross-gender conversion.

It may be concluded that the technique using multivariate quadratic modeling for spectral mapping is useful for voice conversion without requiring a large training data or grouping into acoustic classes. It needs to be further evaluated for voice conversion involving a larger set of speaker pairs and different types of speech material. Its application for voice conversion along with other speech analysis-synthesis platforms needs to be investigated. Its usefulness in voice conversion applications where the speaker pair speaks more than one language may be investigated by training it with parallel data in one language and applying for voice conversion in another.

[blank]

Chapter 6

SUMMARY AND CONCLUSIONS

6.1 Introduction

The spectral parameters are considered to be relatively more important than those related to rhythm and intonation for conveying speaker identity, [19], [20], [22]-[27]. The research objective was to investigate the modification of spectral characteristics for voice conversion by modeling the relationship between the acoustic spaces of the source and the target using a single transformation function.

A novel technique has been proposed based on the hypothesis that a single transformation function applicable to all acoustic classes can be derived using multivariate polynomial modeling. Each parameter for generating the target speech is modeled as a multivariate polynomial function of the parameters of the source speech. The technique has been applied for voice conversion using parallel speech data for training. The set of transformation functions are obtained from the time aligned source and target feature vectors. Voice conversion of the source speech signal is carried out by applying the estimated mapping for modification of spectral characteristics along with pitch and time scaling. Pitch scaling is used to match the range of the pitch in the source speech to that in the target speech. The pitch contour is modified without disturbing the duration of the speech signal. Time scaling is used to approximately match the duration of the source speech to that of the target, using a scaling factor equal to the ratio of the total duration of voiced frames in the target speech to that in the source speech, keeping the pitch contour intact. Evaluation is carried out using objective measures and listening tests.

6.2 Summary of the investigations

A voice conversion system consists of two major building blocks: (i) a speech analysis-synthesis platform and (ii) a system for modification of the parameters in the analysis-synthesis parameters. Harmonic-plus-noise model (HNM) has been used as the analysis-synthesis platform, as it provides high quality speech output with a reasonable number of parameters, and easily permits time and pitch scaling [70], [71].

An implementation of HNM was developed, using GCI detection for pitch-synchronous processing, conversion of the spectral parameters to cepstral coefficients, and estimation of phase function from the estimated harmonic magnitudes. Investigations were carried out to examine the suitability of analysis-synthesis platform for voice conversion. The implementation involving only one set of feature vectors for the voiced frames was found to be well suited for voice conversion. It was established that the transformation of time-varying F_m is not needed and it may be fixed at a value within 4–8 kHz. An introduced jitter of more than 5% in the GCI detection resulted in quality degradation, with a more pronounced effect in utterances with longer vowel segments. The accuracy of the GCI detection method used in HNM implementation with reference to simultaneously recorded electroglottogram was found to be acceptable. Presence of additive white noise in the input speech did not result in a loss in the quality of the synthesized output with respect to that of the input. The output synthesized with estimated phase was intelligible but less natural sounding than the output synthesized with source phase. As difference in the quality is small and the source phase information is lost during modification of the magnitude spectrum, the estimated phase may be used for voice conversion. Lastly, it was seen that the implementation can be used for pitch and time scaling with a wide range of scaling factors for male and female speech. Thus it is suitable for voice conversion across speakers with different pitch ranges and speaking rates.

After establishing the HNM based analysis-synthesis platform and implementing the proposed voice conversion system using multivariate polynomial modeling for spectral modification along with pitch and time scaling, a set of four experiments was carried out for its evaluation using informal listening, objective measures, and subjective listening tests. Conclusions from these experiments are summarized in the next section.

6.3 Conclusions

Estimations of transformation functions for multivariate polynomial models with a degree higher than two were found to involve ill-conditioned matrices. Hence investigations were carried out for univariate linear model (ULM), multivariate linear model (MLM), and multivariate quadratic model (MQM). Multivariate quadratic modeling was also investigated by applying the technique by dividing the feature vectors into classes using k-means algorithm, with separate transformation function estimated for each class. During conversion, the feature vector was first assigned to a class and then converted by using the corresponding transformation function. Transformations using 32 and 64 classes are referred to as MQM32 and MQM64. ULM and MLM involved slightly less computation time as compared to MQM. Listening of converted speech showed that ULM resulted in poor quality speech, while the outputs from MLM and MQM were natural sounding. MQM64 and MQM32 took a bit longer

computation time as compared to MQM, and they resulted in a larger decrease in the distance between the target and transformed speech. But MQM32 and MQM64 outputs had distinct audible distortions affecting the intelligibility and naturalness, indicating that dividing feature vectors into classes adversely affected the transformation. Hence it was concluded that in comparison to ULM, MLM, MQM32, and MQM64, MQM is much better suited for spectral modification.

The conclusions from the experiments involving voice conversion for four speaker pairs (male-male, female-female, male-female, and female-male) can be summarized as the following.

1) During training for estimation of transformation function, removal of redundant feature vectors using a distance threshold could be used for achieving a more uniform sampling of acoustic spaces and improving the estimation of transformation functions. Transformation functions were found to stabilize for training data consisting of ten or more sentences, i.e. with speech data of about a minute.
2) Voice conversion using MQM and GMM (with 64 mixture components) using the same training and test data resulted in almost similar output quality and decrease in target-transformed distance, but MQM needed a much shorter (1/83) computation time for estimation of the transformation function.
3) The parallel training data should have adequate representation of feature vectors corresponding to different acoustic classes. MQM based transformation function can be estimated by using a limited training data if it contains speech segments similar to all speech segments likely to occur during the testing phase.
4) Results of subjective evaluation showed that combination of spectral modification and pitch scaling resulted in transformed speech having good quality and almost the same identity as the target. Averaged across listeners, the scores for identification of transformed speech as the target were 93% for same-gender conversion and 100% for cross-gender conversion.

Thus the investigations have shown that the proposed technique of spectral modification can be used for voice conversion and it may be particularly suitable for applications involving voice conversion with a relatively limited speech data for training.

6.4 Suggestions for future work

Although the proposed voice conversion technique results in perfectly intelligible speech, it results in small audible distortions, particularly in cross-gender conversions, which may be attributed to large pitch scaling factors. Some of these distortions may have also resulted due to occasional errors in GCI detection, discontinuities in the transformed spectra, and

discontinuities in estimated phase. Hence the proposed voice conversion technique needs to be investigated by further modification in the HNM analysis-synthesis platform. One of the possibilities is to use the pitch-synchronous analysis as used in HNM-3 along with fixed-frame synthesis with an overlap-add for reducing the effects of discontinuities. Use of the proposed technique with other analysis-synthesis platforms also needs to be investigated. Some of the audible distortions may have resulted due to uniform time scaling during voiced segments and hence use of non-uniform time scaling methods should be examined.

The technique has to be investigated using a larger number of speaker pairs and different types of speech material. The effect of differences between the source and target speakers with reference to speaking style, jitter, and shimmer needs to be investigated. Its application for speaker pairs speaking two or more languages, with the mapping obtained from the parallel speech data in one language used for voice conversion in another language, may also be investigated. The technique has been applied to voice conversion using parallel speech data for training with time-aligned feature vectors. Its application to non-parallel speech data by establishing correspondence between feature vector clusters of the source and target speech needs to be investigated.

Appendix A

RESULTS OF INVESTIGATIONS USING HNM

Table A.1. Investigation I: Effect of HNM variants. PESQ-MOS test scores, averaged over the utterances in the set, male (M), female (F), and M&F for HNM-1, HNM-2, and HNM-3 based analysis-synthesis (reference: recorded speech). SD: standard deviation.

Set	No. of utterances	HNM-1		HNM-2		HNM-3	
		Mean	SD	Mean	SD	Mean	SD
M	9	2.82	0.09	2.90	0.13	2.91	0.13
F	9	2.96	0.12	3.00	0.13	3.01	0.14
M&F	18	2.89	0.13	2.95	0.14	2.96	0.14

Table A.2. Investigation II: Effect of maximum voiced frequency on synthesized speech using HNM-3 based analysis-synthesis: Mean and SD of PESQ-MOS test scores for synthesized utterances using different values of F_m (reference: recorded speech), averaged over male (M), female (F), M&F sets of utterances.

F_m (kHz)	M		F		M&F	
	Mean	SD	Mean	SD	Mean	SD
1.0	2.15	0.11	1.88	0.17	2.02	0.20
1.5	2.40	0.15	2.30	0.11	2.35	0.13
2.0	2.55	0.16	2.58	0.14	2.56	0.15
2.5	2.71	0.15	2.67	0.17	2.69	0.16
3.0	2.83	0.11	2.84	0.13	2.84	0.12
3.5	2.88	0.12	2.99	0.13	2.94	0.14
4.0	2.90	0.12	3.01	0.12	2.96	0.13
4.5	2.90	0.13	3.00	0.12	2.95	0.13
5.0	2.92	0.13	3.00	0.14	2.96	0.14
5.5	2.90	0.13	3.00	0.14	2.95	0.14
6.0	2.92	0.13	3.00	0.12	2.96	0.13
6.5	2.92	0.14	3.01	0.13	2.97	0.14
7.0	2.90	0.12	3.00	0.13	2.95	0.13
7.5	2.91	0.13	3.01	0.13	2.96	0.14
8.0	2.92	0.13	3.02	0.11	2.97	0.13
Est.	2.91	0.13	3.01	0.14	2.96	0.14

Table A.3. Investigation III: Effect of jitter introduced in GCI estimation: PESQ-MOS test scores for synthesis using different amount of jitter, averaged over male (M), female (F), M&F sets of utterances for HNM-3 based analysis-synthesis (reference: recorded speech). γ = jitter control factor.

γ	Jitter	M		F		M&F	
	(%)	Mean	SD	Mean	SD	Mean	SD
0.00	1.58	2.91	0.13	3.01	0.13	2.96	0.13
0.01	1.66	2.90	0.12	3.01	0.12	2.95	0.13
0.02	1.86	2.89	0.11	2.95	0.12	2.92	0.12
0.03	2.16	2.86	0.11	2.86	0.08	2.86	0.09
0.04	2.59	2.81	0.09	2.74	0.09	2.78	0.10
0.05	2.92	2.78	0.10	2.62	0.07	2.70	0.12
0.06	3.26	2.72	0.11	2.47	0.10	2.60	0.16
0.07	3.61	2.67	0.09	2.36	0.11	2.52	0.19
0.08	3.92	2.62	0.10	2.24	0.12	2.43	0.22
0.09	4.32	2.56	0.10	2.13	0.11	2.34	0.25
0.10	4.72	2.50	0.11	1.98	0.13	2.24	0.29
0.11	4.81	2.45	0.13	1.92	0.14	2.18	0.30
0.12	5.27	2.39	0.13	1.85	0.13	2.12	0.31
0.13	5.48	2.34	0.11	1.76	0.18	2.05	0.33
0.14	5.83	2.31	0.10	1.69	0.13	2.00	0.34
0.15	6.10	2.25	0.15	1.61	0.12	1.93	0.36
0.16	6.21	2.13	0.17	1.48	0.11	1.81	0.36
0.17	6.33	2.19	0.15	1.56	0.15	1.88	0.35
0.18	6.50	2.09	0.16	1.42	0.12	1.75	0.37
0.19	6.69	2.01	0.19	1.42	0.14	1.71	0.34
0.20	6.73	2.09	0.15	1.46	0.12	1.77	0.35

Table A.4. Investigation IV: Effect of input SNR on HNM-3 based analysis-synthesis: PESQ-MOS test scores averaged over male (M), female (F), M&F sets of utterances (reference: recorded speech).

Input	M				F				M&F			
SNR	Noisy		Synth.		Noisy		Synth.		Noisy		Synth.	
(dB)	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$\propto$	4.5	0.00	2.91	0.13	4.5	0.00	3.01	0.14	4.5	0.00	2.96	0.14
20	2.99	0.17	2.59	0.08	2.88	0.29	2.42	0.03	2.94	0.22	2.51	0.11
18	2.84	0.14	2.52	0.09	2.72	0.27	2.36	0.05	2.78	0.20	2.44	0.11
16	2.68	0.13	2.45	0.09	2.57	0.27	2.26	0.06	2.63	0.20	2.36	0.13
14	2.53	0.13	2.37	0.10	2.43	0.27	2.19	0.08	2.48	0.20	2.28	0.13
12	2.40	0.12	2.30	0.10	2.29	0.26	2.11	0.10	2.34	0.19	2.20	0.14
10	2.24	0.11	2.19	0.11	2.16	0.26	2.01	0.14	2.20	0.18	2.10	0.15
8	2.10	0.10	2.10	0.10	2.03	0.25	1.90	0.13	2.06	0.18	2.00	0.15
6	1.95	0.11	1.99	0.12	1.91	0.25	1.81	0.18	1.93	0.17	1.90	0.16
4	1.81	0.11	1.88	0.12	1.79	0.24	1.73	0.17	1.80	0.17	1.81	0.16
2	1.67	0.12	1.77	0.13	1.71	0.25	1.67	0.17	1.69	0.18	1.72	0.15
-0	1.54	0.13	1.65	0.14	1.61	0.24	1.58	0.18	1.57	0.18	1.61	0.15
-2	1.41	0.17	1.55	0.16	1.54	0.26	1.53	0.20	1.47	0.20	1.54	0.16
-4	1.30	0.17	1.45	0.15	1.46	0.23	1.47	0.17	1.38	0.20	1.46	0.14
-6	1.17	0.20	1.35	0.20	1.38	0.24	1.39	0.20	1.28	0.23	1.37	0.18
-8	1.07	0.19	1.26	0.19	1.32	0.22	1.34	0.17	1.19	0.23	1.30	0.17
-10	0.98	0.22	1.17	0.20	1.24	0.22	1.28	0.18	1.11	0.24	1.23	0.18

Table A.5. Investigation V: Effect of phase estimation methods. PESQ-MOS test scores averaged over 9 utterances (3 utterances × 3 speaker pairs) for male and female speech (reference: recorded speech).

Method		Male			Female		
		$p = 14$	$p = 18$	$p = 22$	$p = 14$	$p = 18$	$p = 22$
AS	Mean	2.90	2.92	2.92	2.99	3.03	3.03
	SD	0.13	0.12	0.13	0.14	0.15	0.13
MP	Mean	2.65	2.86	2.86	2.67	2.87	2.87
	SD	0.11	0.10	0.10	0.11	0.10	0.13
SP	Mean	2.80	2.90	2.91	2.80	2.98	2.99
	SD	0.10	0.08	0.08	0.10	0.13	0.14

[blank]

Appendix B

GMM BASED SPECTRAL MAPPING

B.1 Introduction

Gaussian mixture model (GMM) can be used to provide a parametric representation of the mapping from the source acoustic space to the target acoustic space. It has been reported as an efficient system for spectral mapping in voice conversion [20], [56], [94], [224], [233], [234]. It is implemented for a comparison with the proposed spectral mapping based on multivariate polynomial modeling. The model and its implementation for spectral mapping is described in this appendix.

B.2 Gaussian mixture model

For one-dimensional variable x, Gaussian probability density function (PDF) is defined as

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}(x-\mu)^2 / \sigma^2 \right) \tag{B.1}$$

where μ and σ are mean and standard deviation of x. This function is extended to p-dimensional variable $\mathbf{x}$ as

$$g(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\mathbf{\Sigma}|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} (\mathbf{\Sigma})^{-1} (\mathbf{x}-\boldsymbol{\mu})\} \tag{B.2}$$

Here mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$ are $p \times 1$ and $p \times p$ dimensional matrices, respectively [52], [53], [235].

In GMM, a multimodal PDF is approximated by a weighted linear combination of Gaussian functions [56], [234]-[236]. Thus a multidimensional probability density function for vector $\mathbf{x}$ is approximated by a summation of a finite (say m) number of Gaussian components defined as

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^{m} w_k g_k(\mathbf{x}) \tag{B.3}$$

where $p(\mathbf{x}|\boldsymbol{\theta})$ is read as probability density for vector $\mathbf{x}$ corresponding to model $\boldsymbol{\theta}$. Each Gaussian function is assumed to be representing one of the classes of the multimodal random

process. The model $\boldsymbol{\theta}$ represents the collection of parameters of the Gaussian functions and can be given as

$$\boldsymbol{\theta} = \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid k = 1:m\} \tag{B.4}$$

The mixture weights, w_k, satisfy the constraint $\sum_{k=1}^{m} w_k = 1$. Probability that a vector $\mathbf{x}$ belongs to a class c_k is calculated using Bayes' theorem [237]

$$p(c_k \mid \mathbf{x}) = \frac{w_k g_k(\mathbf{x})}{\sum_{i=1}^{m} w_i g_i(\mathbf{x})} \tag{B.5}$$

Substituting (B.2) in (B.5) gives

$$p(c_k \mid \mathbf{x}) = \frac{w_k \left|\boldsymbol{\Sigma}_k\right|^{-1/2} \exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)]}{\sum_{i=1}^{m} w_i \left|\boldsymbol{\Sigma}_i\right|^{-1/2} \exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)]} \tag{B.6}$$

The GMM parameters are iteratively estimated using the expectation maximization (EM) algorithm, with iterations carried out until convergence [53], [55], [71], [119]. Generally, to limit the amount of computation during iteration, an upper limit is set for the number of iterations. Accuracy of the approximation of a PDF using GMM depends upon the size of data, amount of interaction among different dimensions, number of mixture components, initial values of the parameters for iteration, and number of iterations [238]-[240].

In applying GMM on speech feature vectors, improper initialization may result in incorrect parameters, particularly for the speech sounds changing abruptly, such as plosives. The mixture weights of some components may approach zero and the variances may become very large. Generally, three types of initializations are used: random, phone-dependent, and VQ-based. In the first one, the initial estimates are selected randomly. Phone-dependent algorithms use hypothesis of the phone class and an incorrect hypothesis may affect the final estimates. We have used VQ-based initialization employing k-means algorithm [230] with clustering for the initial estimates. From the set of feature vectors, *m* vectors are randomly selected as the initial class centroids. Each vector is assigned to one of the classes based on the criterion of the minimum distance from the class centroids. The mean of all vectors in a class is taken as the new centroid of the class. With these updated centroids, the process of assigning each of the feature vectors to the classes is repeated. The algorithm is assumed to have converged when the means stops changing. The mixture weights are initialized as the ratio of the number of feature vectors in a class to the total number of feature vectors.

Given a set of *N* independent feature vectors [$\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, …, $\mathbf{x}_N$], the log-likelihood for the model $\boldsymbol{\theta}$ [119], [241] may be given by

$$L(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log \sum_{k=1}^{m} w_k g_k(\mathbf{x}_n) \tag{B.7}$$

The estimation starts with an initialization, as described earlier. Each iteration consists of three steps: E-step, M-step, and I-step. In E-step, the conditional probabilities are estimated using (B.6) for each class. In M-step, GMM parameters are updated according to the conditional probabilities estimated in E-step as the following

$$w_k^{j+1} = \frac{1}{N} \sum_{n=1}^{N} p^j(c_k \mid \mathbf{x}_n) \tag{B.8}$$

$$\boldsymbol{\mu}_k^{j+1} = \frac{\sum_{n=1}^{N} p^j(c_k \mid \mathbf{x}_n)\mathbf{x}_n}{\sum_{n=1}^{N} p^j(c_k \mid \mathbf{x}_n)} \tag{B.9}$$

$$\boldsymbol{\Sigma}_k^{j+1} = \frac{\sum_{n=1}^{N} p^j(c_k \mid \mathbf{x}_n)(\mathbf{x}_n - \boldsymbol{\mu}_k^{j+1})(\mathbf{x}_n - \boldsymbol{\mu}_k^{j+1})^{\mathrm{T}}}{\sum_{n=1}^{N} p^j(c_k \mid \mathbf{x}_n)} \tag{B.10}$$

where superscript *j* indicates the values of parameters at *j*th iteration. During I-step, the earlier estimates of the parameters are replaced by the updated ones. These steps are repeated until log-likelihood $L(\boldsymbol{\theta})$ stops increasing. As reported earlier in [241], the convergence in the log-likelihood was found to take place in about 25 iterations. In our implementation, an upper limit for iterations is set as 100.

B.3 Transformation function estimation

Let the sets of the corresponding feature vectors of the source and target speech signals be $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_N]$ and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_N]$, respectively. Let the GMM model of source speech be

$$\boldsymbol{\theta}_x = \{w_{xk}, \boldsymbol{\mu}_{xk}, \boldsymbol{\Sigma}_{xk} \mid k = 1{:}\,m\}$$

and that of target speech be

$$\boldsymbol{\theta}_y = \{w_{yk}, \boldsymbol{\mu}_{yk}, \boldsymbol{\Sigma}_{yk} \mid k = 1{:}\,m\}$$

The mapping from the acoustic space of the source to that of the target, is taken as a linear transformation function

$$F(\mathbf{x}) = \sum_{k=1}^{m} p(c_k \mid \mathbf{x})[\boldsymbol{\nu}_k + \boldsymbol{\Gamma}_k \boldsymbol{\Sigma}_{xk}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{xk})] \tag{B.11}$$

where the vector $\boldsymbol{\nu}_k$ and the matrix $\boldsymbol{\Gamma}_k$ may be determined for minimizing the error [20],

$$\varepsilon = E\left[\, \|\mathbf{y} - F(\mathbf{x})\|^2 \,\right] \tag{B.12}$$

over the training data. Kain and Macon [94] have shown that a joint GMM can be fitted on the set $\mathbf{Z} = [\mathbf{X}^T\ \mathbf{Y}^T]^T$ as

$$\boldsymbol{\theta} = \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid k = 1:m\}$$

The mean vector and covariance matrix of this model are given as

$$\boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_{xk} \\ \nu_k \end{bmatrix} \tag{B.13}$$

$$\boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{\Sigma}_{xk} & \boldsymbol{\Gamma}_k^{\mathbf{T}} \\ \boldsymbol{\Gamma}_k & \boldsymbol{\Sigma}_{yk} \end{bmatrix} \tag{B.14}$$

Thus the vector $\boldsymbol{\nu}_k$ and the matrix $\boldsymbol{\Gamma}_k$ are obtained from $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$.

B.4 Transformation of source speech

For estimating the GMM-based transformation function, the speech material and the processing steps of segmentation, HNM analysis, spectral parameters conversion to MFCC-based feature vectors, DTW alignment, and elimination of redundant feature vectors are the same as those used for multivariate polynomial modeling, as described in Chapter 4. Two transformation functions, each for voiced and unvoiced segments, are estimated separately. The scheme for transformation of source speech is also the same as the one used with multivariate polynomial modeling.

Appendix C

EVALUATION METHODS

C.1 Introduction

The methods used for the evaluation of output speech can be broadly classified as subjective and objective. Subjective methods involve listening tests using human subjects. They provide a real assessment of the quality, but are expensive and time consuming. The objective methods are based on computation. They ensure uniformity in the evaluation, but may not indicate the real quality [242]-[249]. Most of the methods were devised for evaluating speech communication systems, but they are also used for the evaluation of the voice conversion systems.

C.2 Subjective evaluation

The results of the subjective evaluation may get affected by the test conditions, and hence these have to be standardized and consistently followed. The subjects should be adequately familiarized with the reference quality before the test. The subjective tests may be grouped in three categories: intelligibility, quality, and identity.

Intelligibility tests

It measures the percentage of correctly received stimuli as recognition scores. Use of nonsense monosyllables as the stimuli gives the articulation score and use of meaningful words as the stimuli gives word intelligibility score. Different kinds of word lists have been constructed, considering phonetic balance (PB), word length, stress position, word importance, etc. Examples of word lists are the rhymed syllable list, two-syllable Spondee word list, and monosyllable PB word list [250]-[252].

Quality tests

Quality of the phrases is generally evaluated by mean opinion score (MOS), degradation category rating (DCR), and preference tests. In MOS test or absolute category test, the subject rates the quality of the speech stimuli on 1-5 scale (1: bad, 2: poor, 3: fair, 4: good, 5:

excellent). The stimuli are presented in a randomized order, with three to five presentations of each stimulus. The average score calculated across stimuli and subjects is known as the mean opinion score (MOS) [148], [253], [254]. In one variant of this test, the scale used is 0-4 instead of 1-5 (0: bad, 1: poor, 2: fair, 3: good, 4: excellent) [255], [256]. The test gives an assessment based on all the parameters affecting the quality. It is easy to conduct and does not need trained listeners, but its sensitivity for high quality speech is low.

In DCR or DMOS test, the quality is rated with respect to a reference phrase [257]. The stimuli in the form of phrases are presented in a randomized order as pairs (A-B) or repeated pairs (A-B-A-B) where A is a the reference stimulus and B is the test stimulus. The reference serves as an anchor for the subject's judgments. The subject rates the quality of the test stimulus on a 1-5 scale (1: degradation is very annoying, 2: degradation is annoying, 3: degradation is slightly annoying, 4: degradation is audible but not annoying, 5: degradation is inaudible). A few null pairs (A-A-A-A) can also be included in the test sequences to assess the quality of anchoring of the listeners' judgments [126]. As the test uses an annoyance scale and a high quality reference before each judgment, it is considered as suitable for evaluating good quality speech [258].

In preference test, also known as the paired comparison test or the AB test, pairs of stimuli in the form of phrases (A and B) are presented to the subject. The subject responds after each presentation by choosing the better signal. The percentage of presentations in which the test signal is the preferred signal is the preference score. The main advantage of the test is that it involves a binary judgment and hence it gives precise results. However, it is difficult to judge between a pair of speech signals with different types of degradation.

Identity tests

Opinion test and XAB test are conducted for quantifying the identity of the speaker. In opinion test, the subject rates the similarity of each pair of speakers on a 0-9 scale (0: identical and 9: very different) [37] or 1-5 scale (1: similar, 2: a little similar, 3: no-judgment, 4: a little dissimilar, 5: dissimilar) [20], [57], [91], [140], [148]. In a variant of this test, the subject has to decide if the voices come from different speakers on 0-4 scale (0: bad, 4: excellent) [255], [256]. In XAB test, for evaluating voice conversion, a set of triads of phrases are presented to the listeners. Stimulus X may be the source, target, or the modified speech stimulus. Stimuli A and B are either from the target or the source speaker. Speakers A and B use the same sentence which, in general, may be different from the sentence uttered by X. The subject selects either A or B as being more similar to X [20], [57], [259].

C.3 Objective evaluation

The objective methods for quality evaluation estimate the difference between the test and the reference speech signals using computational methods. As the difference is affected by the degree of time-alignment of the corresponding frames in the test and the reference signals, the two signals should be accurately aligned.

Spectral distance

It has been used with many variants, such as magnitude spectral distance, log spectral distance, or distance between feature vectors (e.g. MFCCs, log area ratios) [34], [116], [260]-[263]. Frame-wise distance between the test spectrum S_t and reference spectrum S_r is given as

$$d(t,r)=\left[\frac{1}{K}\sum_{k=1}^{K}\left|S_t(k)-S_r(k)\right|^2\right]^{1/2} \tag{C.1}$$

Frame-wise distance measures may be combined to obtain average RMS spectral distance. For example in [18], RMS magnitude spectral distance is estimated using N frames

$$D(t,r)=\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{K}\sum_{k=1}^{K}\left|S_{t,i}(k)-S_{r,i}(k)\right|^2\right]^{1/2} \tag{C.2}$$

In perceptually weighted distance (PWD), the spectra are normalized for equalizing areas under log spectra and the distance measure is calculated as

$$d(t,r)=\frac{\sum_{k=1}^{K}W(k)\left|S_t(k)-S_r(k)\right|}{\sum_{k=1}^{K}W(k)S_r(k)} \tag{C.3}$$

where S_r and S_t are the normalized power spectra of the reference and test signals, respectively and $W(k)$ is a weighting function for perceived loudness of different spectral components. It has been reported that the measure produces results consistent with published subjective perceptual results on formant frequency difference limens [116].

For log spectral distance (LSD), the distance over a frame is calculated as

$$d(t,r)=\left[\frac{1}{K}\sum_{k=1}^{K}\left|\log\left(S_t(k)\right)-\log\left(S_r(k)\right)\right|^2\right]^{1/2} \tag{C.4}$$

An average RMS log spectral distance [254], [138] over frames of an utterance is given by

$$D(t,r)=\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{K}\sum_{k=1}^{K}\left|\log\left(S_{t,i}(k)\right)-\log\left(S_{r,i}(k)\right)\right|^2\right]^{1/2} \tag{C.5}$$

A multiplication by 20 can be used to express the distance in dB. The results of this measure are not as consistent as that of the PWD based measure [116]. The average LSD based

measure between two identical utterances spoken by the same speaker can vary by 5−10 dB, while the distance between two different speakers would normally be 13−20 dB [36].

The log area ratios (LAR) measure estimates the absolute deviation in the log area ratios which are uniquely related to the reflection coefficients obtained from LPC analysis, and possess flat or uniform spectral sensitivity [264]. Let $g_t(n)$ and $g_r(n)$ be the log area ratios of the target *t* and reference speaker *r* for the section *n* of the vocal tract, the distance between them is defined as

$$d(t,r) = \frac{1}{p}\sum_{n=1}^{p}\left|g_t(n) - g_r(n)\right| \tag{C.6}$$

An objective measure based on cepstrum coefficients [126], [265], [266] has been defined as

$$d(t,r) = \frac{10}{\log 10}\sqrt{2\sum_{n=1}^{p}\left[c_t(n) - c_r(n)\right]^2} \tag{C.7}$$

In this measure, the cepstral distance is limited in the range [0, 10] to minimize the number of outliers. Some of the other methods for defining cepstral distances [2], [20], [113], [142], are as the following

$$d(t,r) = 2\sum_{n=1}^{p}\left[c_t(n) - c_r(n)\right]^2 \tag{C.8}$$

$$d(t,r) = \frac{1}{p}\sum_{n=1}^{p}\left[c_t(n) - c_r(n)\right]^2 \tag{C.9}$$

By defining a cepstral distance vector $\mathbf{d}_i = \left[\, d_i(1) \quad d_i(2) \quad d_i(3) \quad \cdots \quad d_i(p)\right]$ for frame *i* with $d_i(n) = c_t(n) - c_r(n)$, the cepstral Mahalanobis distance is given as [267]-[271]

$$d(t,r) = \sqrt{\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{d}_i} \tag{C.10}$$

where $\boldsymbol{\Sigma}$ the covariance matrix calculated over all the frames. The distances over the frames can be used to obtain an average distance

$$D(t,r) = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{d}_i} \tag{C.11}$$

Mahalanobis distance has been reported to be an efficient measure for multidimensional pattern comparisons [268]-[270], [272], and has been often used for distance in parametric space in speech research [271], [273].

In [274], the distance values were averaged across frames for getting a single speech quality score using an energy dependent weighting function. This weighting function was estimated by assuming that the frame errors in low energy regions have a smaller influence on quality than those in high energy regions. With $W'(i)$ as the weighting function denoting the

speech signal frame energy per sample in dB computed over 2 ms, the distance was computed as *r*th mean over *N* frames by using

$$D(t,r)=\left[\frac{\sum_{i=1}^{N} W'(i)\left|d_i(t,r)\right|^r}{\sum_{i=1}^{N} W'(i)}\right]^{1/r} \tag{C.12}$$

As the perceived quality may be influenced by the presence of even a small number of frames with large errors such as those that are perceived as pops or glitches, a simple arithmetic average computed for r=1 may not correctly give the quality measure. Use of a larger value of r emphasizes the effect of frames having large errors. A composite score was also calculated by combining the simple arithmetic mean ($D_a(t,r)$) and average of the top 10% of the frame errors ($D_b(t,r)$) estimated from (C.12) with r=1.

$$D(t,r)=D_a(t,r)+k_1 e^{-\alpha} D_b(t,r) \tag{C.13}$$

where k_1 is a constant, and α is the skewness factor estimated from statistics of the frame distances [274]

$$\alpha=\frac{1}{N}\sum_{i=1}^{N}\left[d_i(t,s)-\mu_d\right]^3/\sigma_d^3$$

Three distance measures, namely PWD, LSD, and LAR have been compared in [274]. The results of LAR using (C.12) and PWD using (C.2) were found to be inferior to those of LSD using (C.4) and LAR using (C.6). The composite score calculated using (C.13) was reported to be a superior method of combining frame errors.

As the absolute distance between target and the transformed source does not indicate the perceptual distance of the transformed speech to the actual target speech, the relative distance may be estimated for evaluating the performance of the speaker transformation system [61] using

$$D(t,t')=\frac{\sum_{i=1}^{N} d_i(t,t')}{\sum_{i=1}^{N} d_i(t,s)} \tag{C.14}$$

where t, t', and s denote the speech signals of the target, transformed source, and source speakers, respectively. In order to have a normalized error across different speaker combinations, Kain and Macon [24] defined a performance index (PI) as

$$\text{PI}=1-\frac{D(t,t')}{D(t,s)} \tag{C.15}$$

Log likelihood ratio

Log likelihood ratio (LLR) or objective XAB uses a speaker recognition system to measure the closeness of a given speech to a given speaker. If the system is trained for two speakers 'A' and 'B', the system results in negative value when the input speech is closer to that of speaker 'A' and a positive value if it is closer to that of speaker 'B' [35], [50], [275]. For a given input utterance X, the LLR of the speaker 'B' to the speaker 'A' is defined as

$$\theta = \log \frac{p\left(\mathrm{X}|\lambda_{\mathrm{B}}\right)}{p\left(\mathrm{X}|\lambda_{\mathrm{A}}\right)} \tag{C.16}$$

where λ_{B} and λ_{A} are the speaker models for 'A' and 'B', respectively. Speaker models may be built using several techniques. In GMM-UBM technique [235], [254] a universal background model (UBM) is estimated from the training involving hundreds of speakers and then speaker-specific models λ_{B} and λ_{A} are built using maximum a posteriori (MAP) adaptation. For performing an objective XAB test, the given stimulus X is evaluated against the trained UBM, and the top scoring mixture components are found. The degree of similarity between the input X and the predefined models of the speaker 'A' and 'B' is computed only for the top scoring mixture components. Finally, the log-likelihood ratio is computed as

$$\theta = \frac{\log\left(p\left(\mathrm{X}|\lambda_{\mathrm{B}}\right)\right) - \log\left(p\left(\mathrm{X}|\lambda_{\mathrm{UBM}}\right)\right)}{\log\left(p\left(\mathrm{X}|\lambda_{\mathrm{A}}\right)\right) - \log\left(p\left(\mathrm{X}|\lambda_{\mathrm{UBM}}\right)\right)} \tag{C.17}$$

The statistics of the scores for a large number of utterances X allows improving the robustness of the performance estimation by calculating

$$D_{\mathrm{NORM}} = \frac{\mu_{\mathrm{B}} - \mu_{\mathrm{X}}}{\mu_{\mathrm{B}} - \mu_{\mathrm{A}}} \tag{C.18}$$

where μ_{A}, μ_{B}, and μ_{X} are the mean values of the distributions for the scores of the speaker 'A', speaker 'B', and the input utterances referred as X, respectively.

PESQ

Perceptual evaluation of speech quality (PESQ) is a procedure for objective evaluation of speech quality using auditory mechanism of the brain and has been found to perform better than the other objective measures [247], [248]. It is a narrow-band speech quality assessment and it gives accurate predictions of subjective quality in a very wide range of conditions, including those with background noise, analogue filtering, and variable delay. It has been incorporated as the ITU-T P.862 Recommendation [231]. For estimating the PESQ score, both test and reference signals are adjusted to a standard listening level. The signals are time aligned assuming the delay introduced by the transmission system as piecewise constant. Frame-by-frame delays are estimated using envelope and fine correlation histogram-based

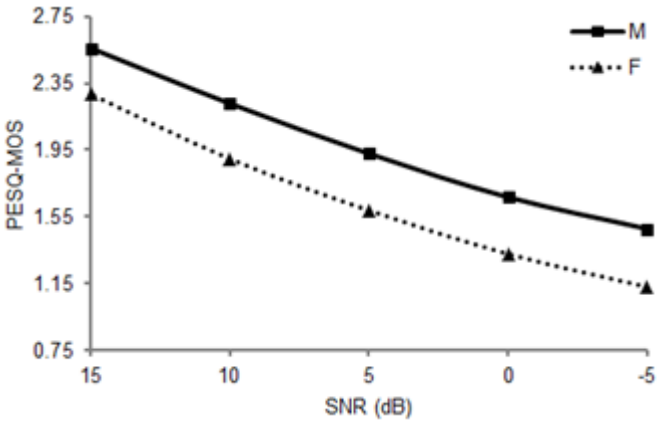

Fig. C.1 Effect of white noise on the PESQ-MOS on male (M) and female (F) speech signal, adapted from Fig. 2 in [276].

delay identification. The perceived loudness in each time-frequency cell is estimated from the Bark spectrum. The frame-by-frame distance between the two input signals is estimated by disturbance processing and cognitive modeling. The distance is multiplied by a weight inversely proportional to the instantaneous energy of the reference signal, raised to the power 0.04, for giving slightly greater emphasis on weak sections. It is multiplied by another weight which accounts for the effect of short-term memory in listening. Generally, the signals of 5−8 s duration are used in the PESQ evaluation. For signals longer than 16 s, the weightage to the frame disturbance is reduced linearly from 1.0 at 16 s to 0.5 at 60 s. Finally, a non-linear averaging representing cognitive modeling is used to generate the score indicating subjective mean opinion score. Maximum value of the score, for a signal being compared with itself, is 4.5. Ma *et al.* [276] reported the effect of white noise on the PESQ-MOS on speech signals from a male and a female speaker (duration: 8 s, sampling: 8 kHz). As shown in Fig. C.1, the score decreased sharply from 4.5 to 2.56, and 4.5 to 2.28, for male and female speech respectively, indicating that the measure is very sensitive to low levels of additive noise. Decrease in the score for SNR value lower than 15 dB is gradual.

C.4 Summary

Several communally used subjective and objective evaluation methods have been briefly described. Informal listening tests showed that voice conversion resulted in perfectly intelligible speech, leading to the conclusion that test for intelligibility are not needed. Therefore listening tests for the evaluation of voice conversion were carried out for two measures: MOS for the quality of the converted speech and XAB for the identity of the converted speech. Objective measures PESQ-MOS and cepstral-based Mahalanobis distance were used in the intermediate stages of investigations.

[blank]

Appendix D

INSTRUCTIONS FOR XAB-MOS TEST

The listening tests are being conducted for research related to voice conversion. The test involves presentations of sets of three speech sentences marked as X, A, and B and your responses to the presentations. In each set, the three speech sounds correspond to the same sentence but may be from different speakers. The first task is to match the voice of X as being close to that of A or B. The second task is to assess the quality of X on 1-5 scale: 1- bad, 2 - poor, 3 - fair, 4 - good, 5 - excellent). For response, quality of A and B may be considered as excellent.

The test will be conducted using a computer for presenting the test sounds as well as for noting your responses. You will have a trial test to become familiar with the sounds and the method. Be relaxed and attentive throughout the test.

1. The speech will be presented through a speaker connected to the computer and the level of the sounds will be adjusted to the most comfortable level for you. You will be giving your responses by clicking the mouse on the buttons displayed on the screen.
2. The screen will have the following buttons
    a) Play X, Play A, and Play B: Press these buttons to listen to speech sounds X, A, and B. After carefully listening to the sounds, you have to match the speaker identity of speech X either to that of A or B. You can listen to each of the three sounds more than once. Give your response by clicking on "Identity" button for the speaker identity and on "Quality" for the quality of the speech presented as X.
    b) Identity: When you press this button, a vertical list will appear with two options labeled as A and B. If the speaker identity of the speech X is closer to that of A, then click on option A otherwise click on option B.
    c) Quality: When you press this button, a vertical list will appear with five options corresponding to the quality of the speech on 1-5 scale (1: bad, 2: poor, 3: fair, 4: good, 5: excellent). Click on the option corresponding to the quality of speech X.
    d) Next: After giving your responses for identity and quality, press this button for the presentation of the next set of sounds.

3. The test will be continued until all the sets of sounds have been presented and your responses noted. At the end of the test, “Test is over” will be displayed on the screen.

REFERENCES

[1] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 5, pp. 954-964, 2010.

[2] A. Mouchtaris, J. Van der Spiegel, and P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 3, pp. 952- 963, 2006.

[3] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 5, pp. 944-953, 2010.

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

[5] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 5, pp. 922-931, 2010.

[6] A. Dustor, "Voice verification based on nonlinear Ho-Kashyap classifier," in *Proc. IEEE Int. Conf. Computational Technologies Elect. and Electron. Eng.*, Novosibirsk, Russia, 2008, pp. 296-300.

[7] S. Barua, "Authentication of cellular users through voice verification," in *Proc. IEEE Int. Conf. Syst., Manage. And Cybern.*, Nashville, TN, 2000, pp. 420 - 425.

[8] A. Mouchtaris, J. V Spiegel, P. Mueller, and P. Tsakalides, "A spectral conversion approach to feature denoising and speech enhancement," in *Proc. 9th EuroSpeech*, Lisbon, 2005, pp. 2057-2060.

[9] K. Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, Istanbul, 2000, pp. 1843-1846.

[10] K. Nakamura, T. Toda, Y. Nakajima, H. Saruwatari, and K. Shikano, "Evaluation of

speaking-aid system with voice conversion for laryngectomees toward its use in practical environments," in *Proc. 9th Annu. Conf. Int. Speech Commun. Assoc.*, Brisbane, Australia, 2008, pp. 2209-2212.

[11] K. Tokuda, T. Masuko, J. Hiroi, T. Kobayashi, and T. Kitamura, "A very low bit rate speech coder using HMM-based speech recognition / synthesis," in *Proc. IEEE Int. Conf. Acoustic, Speech Signal Process.*, Seattle, 1998, pp. 609-612.

[12] M. Mashimo, T. Toda, H. Kawanami, K. Shikano, H. Kashioka, and N. Campbell, "Evaluation of cross-language voice conversion evaluation using bilingual databases," in *Proc. 7th Int. Conf. on Spoken Language Process.*, Denver, Col., 2002, pp. 293-296.

[13] Y. Sato, "Voice quality conversion using interactive evolution of prosodic control," *Appl. Soft Computing*, vol. 5, no. 2, pp. 181-192, 2005.

[14] C. H. Wu, C. C. Hsia, T. H. Liu, and J. F. Wang, "Voice conversion using duration-embedded Bi-HMMs for expressive speech synthesis," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 4, pp. 1109-1116, 2006.

[15] M. Eichner, M. Wolff, and R. Hoffmann, "Voice characteristics conversion for TTS using reverse VTLN," in *Proc. IEEE Int. Conf. Acoust,. Speech, Signal Process.*, Montreal, 2004, pp. 17-20.

[16] J. Nurminen, V. Popa, J. Tian, and I. Kiss, "A parametric approach for voice conversion," in *Proc. TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, 2006.

[17] A. Pribilova and J. Pribil, "Non-linear frequency scale mapping for voice conversion in text-to-speech system with cepstral description," *Speech Commun.*, vol. 48, no. 12, pp. 1691-1703, 2006.

[18] W. S. Percybrooks and E. Moore, "Voice conversion with linear prediction residual estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Las Vegas, NV, 2008, pp. 4673-4676.

[19] D. G. Childers, B. Yegnanarayana, and W. Ke, "Voice conversion: factors responsible for quality," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Tampa, Florida, 1985, pp.748-751.

[20] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Process.*, vol. 6, no. 2, pp. 131-142, 1998.

[21] D. Sundermann, A. Bonafonte, H. Ney, and H. Hoege, "A study on residual prediction techniques for voice conversion," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, 2005, pp. 512-516.

[22] H. Matsumoto, S. Hiki, T. Sone, and T. Nimura, "Multidimensional representation of

personal quality of vowels and its acoustical correlates," *IEEE Trans. Audio Electroacoust.*, vol. 21, no.5, pp. 428-436, 1973.

[23] D. G. Childers, K. Wu, D. M. Hicks, and B. Yegnanarayana, "Voice conversion," *Speech Commun.*, vol. 8, no. 2, pp. 147-158, 1989.

[24] A. Kain and M. W. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Salt Lake City, UT, 2001, pp. 813-816.

[25] K. Itoh, "Perceptual analysis of speaker identity," in *Speech Science and Technology*, S. Saito, Ed. Tokyo: Ohmsha, 1992, pp. 133-145.

[26] W. Percybrooks and E. Moore, "New algorithm for LPC residual estimation from LSF vectors for a VC system," in *Proc. EuroSpeech*, Antwerp, Belgium, 2007, pp. 1977-1980.

[27] J. Sun, B. Dai, J. Zhang, and Y. Xie, "Modeling glottal source for high quality voice conversion," in *Proc. 6th World Congress Intelligent Control Automation*, Dalian, China, 2006, pp. 9459-9462.

[28] H. Mizuno and M. Abe, "Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectral tilt," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Adelaide, Australia, 1994, pp. I/469 -I/472.

[29] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," *Speech Commun.*, vol. 16, no. 2, pp. 207-216, 1995.

[30] M. Savic and I. H. Nam, "Voice personality transformation," *Digital Signal Process.*, vol. 4, no. 2, pp. 107-110, 1991.

[31] K. S. Lee, W. Doh, and D. H. Youn, "Voice conversion using a low dimensional vector mapping," *IEICE Trans. Inf. Syst.*, vol. E85-D, no.8, pp. 1297-1305, 2002.

[32] T. Galas and X. Rodet, "An improved cepstral method for deconvolution of source filter systems with discrete spectra application to musical sound signals," in *Proc. Int. Comput. Music Conf.*, Glasgow, 1990, pp. 82-84.

[33] S. Imai and Y. Abe, "Spectral envelope extraction by improved cepstral method," *Electron. Commun.*, vol. 62, no. 4, pp. 10-17, 1979.

[34] K. S. Lee, D. H. Youn, and I. W. Cha, "A new voice personality transformation based on both linear and nonlinear prediction analysis," in *Proc. Int. Conf. Spoken Lang. Process.*, Philadelphia, PA, 1996, pp. 1401-1404.

[35] L. M. Arslan, "Speaker transformation algorithm using codebooks (STASC)," *Speech Commun.*, vol. 28, no. 3, pp. 211-226, 1999.

[36] H. Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp.

1301-1312, 2006.

[37] O. Turk and L. M. Arslan, “Robust processing techniques for voice conversion,” *Comput. Speech Language*, vol. 20, no.4, pp. 441-467, 2006.

[38] A. Mouchtaris, S. S. Narayanan, and C. Kyriakakis, “Multichannel audio synthesis by subband-based spectral conversion and parameter adaptation,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 263-274, 2005.

[39] K. Shikano, K. Lee, and R. Reddy, “Speaker adaptation through vector quantization,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Tokyo, 1986, pp. 2643-2646.

[40] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “Voice conversion through vector quantization,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, New York, 1988, pp. 655–658.

[41] M. Abe, “A segment-based approach to voice conversion,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Toronto, 1991, pp. 765-768.

[42] Mouchtaris, Y. Agiomyrgiannakis, and Y. Stylianou, “Conditional vector quantization for voice conversion,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Honolulu, HI, 2007, pp. 505-508.

[43] A. Rinscheid, “Voice conversion based on topological feature maps and time variant filtering,” in *Proc. Int. Conf. Spoken Language*, Philadelphia, PA, 1996, pp. 1445-1448.

[44] G. Zuo and W. Liu, “Genetic algorithm based RBF neural network for voice conversion,” in *Proc. 5th World Congr. Intelligent Control and Automation*, Hangzhou, China, 2004, pp. 4215-4218.

[45] Z. H. Jian and Z. Yang, “Voice conversion without parallel speech corpus based on mixtures of linear transform,” in *Proc. Int. Conf. Wireless Commun., Networking and Mobile Computing*, Shanghai, 2007, pp. 2825-2828.

[46] R. M. de Freitas, Y. E. Shimabukuro, and R. R. Rosa, "Wavelets transform and linear spectral mixture model applied to MODIS time series for land cover change analysis,” in *Proc. IEEE Int. Symp. Geoscience and Remote Sensing*, Barcelona, Spain, 2007, pp. 1951-1954.

[47] K. M. Ozonat and R. M. Gray, "Gaussian mixture image classification for the linear image transforms,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, 18-23 March, 2005, vol. 5, pp. v/337- v/340.

[48] S. Furui, “Research on individuality features in speech waves and automatic speaker recognition techniques,” *Speech Commun.*, vol. 5, no. 2, pp. 183-197, 1986.

[49] M. L. Arslan and D. Talkin, “Speaker transformation using sentence HMM based alignments and detailed prosody modification,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Seattle, WA, 1998, pp. 289-292.

[50] O. Salor, M. Demirekler, B. Pellom, “A system for voice conversion based on adaptive filtering and line spectral frequency distance optimization for text-to-speech synthesis,” in *Proc. EuroSpeech*, Geneva, Switzerland, 2003, pp. 2417-2420.

[51] Y. Stylianou and O. Cappe, “A system for voice conversion based on probabilistic classification and a harmonic plus noise model,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Seattle, WA, 1998, pp. 281-284.

[52] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[53] D. Reynolds and R. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” *IEEE Trans. Speech and Audio Process.*, vol. 3, no. 1, pp. 72-83, 1995.

[54] R. C. Rose and D. A. Reynolds, “Text independent speaker identification using automatic acoustic segmentation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Albuquerque, NM, 1990, pp. 293-296.

[55] B. L. Tseng, F. K. Soong, A. E. Rosenberg, “Continuous probabilistic acoustic map for speaker recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, San Francisco, CA, 1992, pp. 161-164.

[56] P. Zolfaghari and A. J. Robinson, “Formant analysis using mixtures of Gaussians,” in *Proc. Int. Conf. Spoken Language*, Philadelphia, PA, 1996, pp. 1229-1232.

[57] H. Valbret, E. Moulines, and J. P. Tubach, “Voice transformation using PSOLA technique,” *Speech Commun.*, vol. 11, no. 2-3, pp. 175-187, 1992.

[58] D. Sundermann, H. Ney, and H. Hoge, “VTLN-based cross-language voice conversion,” in *Proc. IEEE Workshop Automat. Speech Recognition and Understanding*, St. Thomas, VI, 2003, pp. 676-681.

[59] T. Toda, H. Saruwatari, K. Shikano, “Speaker transformation algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Salt Lake City, UT, 2001, pp. 841-844.

[60] J. Slifka and T. R. Anderson, “Speaker modification with LPC pole analysis,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Detroit, MI, 1995, pp. 644-647.

[61] N. Iwahashi and Y. Sagisaka, “Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks,” *Speech Commun.*, vol. 16, no. 2, pp.139-151, 1995.

[62] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Speaker interpolation in HMM-based speech synthesis system,” in *Proc. EuroSpeech*, Rhodes, Greece, 1997, pp.2523-2526.

[63] M. C. Orhan and C. Demiroglu, "HMM-based text to speech system with speaker interpolation," in *Proc. IEEE Conf. Signal Process. Commun. Applicat.*, Kemer, Antalya, Turkey, 2011, pp.781-784.

[64] M. Hashimoto and N. Higuchi, “Training data selection for speaker transformation using speaker selection and vector field smoothening,” in *Proc. Int. Conf. Spoken Language*, Philadelphia, PA, 1996, pp. 1397-1400.

[65] J. I. Takahashi and S. Sagayama, "Vector-field-smoothed Bayesian learning for incremental speaker adaptation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Detroit, MI, 1995, vol. 1, pp. 696-699.

[66] M. Tonomura, T. Kosaka, and S. Matsunaga, "Speaker adaptation based on transfer vector field smoothing using maximum a posteriori probability estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Detroit, Michigan, 1995, vol. 1, pp. 688-691.

[67] M. Hashimoto and N. Higuchi, "Training data selection for voice conversion using speaker selection and vector field smoothing," in *Proc. Int. Conf. Spoken Language*, Philadelphia, PA, 1996, vol. 3, pp. 1397-1400.

[68] A. Rinscheid, "Voice conversion based on topological feature maps and time-variant filtering," in *Proc. Int. Conf. Spoken Language*, Philadelphia, PA, 1996, vol. 3, pp. 1445-1448.

[69] S. Olmos, J. Garcia, P. Laguna, and R. Jane, "Truncated orthogonal expansions of recurrent signals: equivalence to a periodic time-variant filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Seattle, WA, 1998, vol. 3, pp. 1709-1712.

[70] B. Saleh and N. Subotic, "Time-variant filtering of signals in the mixed time frequency domain," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 6, pp. 1479-1485, 1985.

[71] J. Laroche, Y. Stylianou, and E. Moulines, “HNS: Speech modification based on a harmonic + noise model,” in *Proc. Int. Conf. on Acoust., Speech, Signal Process.*, Minneapolis, MN, 1993, pp. 550–553.

[72] P. K. Lehana and P. C. Pandey, “Harmonic plus noise model based speech synthesis in Hindi and pitch modification,” in *Proc. 18th Int. Congr. Acoust.*, Kyoto, Japan, 2004, pp. 3333-3336.

[73] H. Kuwabara and Y. Sagisaka, “Acoustic characteristics of speaker individuality control and conversion,” *Speech Commun.*, vol. 16, no. 2, pp. 165-173, 1995.

[74] H. Kasuya, Y. Kobayashi, and T. Kobayashi, “Characteristics of pitch period and amplitude perturbations in pathological voice,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Boston, MA, 1983, pp. 1372-1375.

[75] H. Kasuya, K. Masubuchi, S. Ebihara, and H. Yoshida, “Preliminary experiments on

voice screening," *J. Phonetics*, vol. 14, no. 3-4, pp. 463-468, 1986.

[76] H. Kasuya, S. Ogawa, Y. Kikuchi, and S. Ebihara, "An acoustic analysis of pathological voice and its application to the evaluation of laryngeal pathology," *Speech Commun.*, vol. 5, no. 2, pp. 171-181, 1986.

[77] H. Muta, T. Muraoka, K. Wagatsuma, H. Fukuda, E. Takayama, T. Fujioka, and S. Kanou, "Analysis of hoarse voices using the LPC method," in *Laryngeal Function in Phonation and Respiration*, T. Bear, C. Sasaki, and K. Harris, Eds. Philadelphia, PA: Lippincott Williams Wilkins, 1987.

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

[79] G. Fant, "Some problems in voice source analysis," *Speech Commun.*, vol. 13, no. 1-2, pp. 7-22, 1993.

[80] I. R. Murray and J. L. Arnott, "Towards the simulation of emotion in synthetic speech: a review of the literature of human vocal emotion," *J. Acoust. Soc. Amer.*, vol. 93, no. 2, pp. 1097-1108, 1993.

[81] I. Karlsson, "Glottal waveform parameters for different speaker types," in *Proc. 7th FASE Symp.*, Edinburgh, 1988, pp. 22-26.

[82] A. L. Lalwani and D. G. Childers, "Modeling vocal disorders via formant synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Toronto, 1991, pp. 505-508.

[83] J. Schoentgen and R. de Guchteneere, "Time series analysis of jitter," *J. Phonetics*, vol. 23, pp. 189-201, 1995.

[84] I. Karlsson, "Modelling voice variations in female speech synthesis," *Speech Commun.*, vol. 11, no. 4-5, pp. 491-495, 1992.

[85] M. Bavegard, G. Fant, J. Gauffin, and J. Liljencrants, "Vocal tract sweeptone data and model simulations of vowels, laterals and nasals," Speech Transmission Laboratory, Royal Inst. of Tech. Stockhom, Sweden, *Quart. Status Rep.*, vol. 34, no. 4, pp. 43-76, 1993.

[86] P. R. Cook, "SPASM, a real time vocal tract physical model controller, and singer, the companion software synthesis system," *Comput. Music J.*, vol. 17, no. 1, pp. 30-44, 1993.

[87] P. H. Milenkovic, "Voice source model for continuous control of pitch period," *J. Acoust. Soc. Amer.*, vol. 93, no. 2, pp. 1087-1096, 1993.

[88] D. G. Childers, *Speech Processing and Synthesis Toolboxes*. New York: Wiley, 1999.

[89] D. G. Childers and C. K. Lee, "Voice quality factors analysis, synthesis and perception," *J. Acoust. Soc. Amer.*, vol. 90, no. 5, pp. 2394-2410, 1991.

[90] Y. Q. Gao and Z. Yang, "Probabilistic approach for speaker transformation," in *Proc. IEEE Int. Conf. Wireless Commun. Netw. Mobile Computing*, Shanghai, 2007, pp. 2845-2848.

[91] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Perez, and Y. Stylianou, "Towards a voice conversion system based on frame selection," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Process.*, Honolulu, HI, 2007, pp. IV.513-IV.516.

[92] A. Mouchtaris, J. V Spiegel, and P. Mueller, "Nonparallel training for voice conversion by maximum likelihood constrained adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Montreal, 2004, pp. 1-4.

[93] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, "Text-independent speaker transformation based on unit selection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Toulouse, France, 2006, pp. I.81-I.84.

[94] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Seattle, WA, 1998, pp. 285-288.

[95] L. Rabiner and B. H. Juang, *Fundaments of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall, 1999.

[96] C. Myers, L. R. Rabiner, and A. E. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 6, pp. 623-635, 1980.

[97] F. Villavicencio, A. Robel, and X. Rodet, "Extending efficient spectral envelope modeling to mel-frequency based representation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, 2008, pp. 1625-1628.

[98] X. Zhou, D. G. Romero, R. Duraiswami, C. E. Wilson, and S. Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition," in *Proc. IEEE Workshop Automat. Speech Recognition and Understanding*, 11-15 Dec., 2011, pp. 559-564.

[99] E. Helander, J. Nurminen, and M. Gabbouj, "LSF mapping for voice conversion with very small training sets," in *Proc. Int. Conf. on Acoust., Speech, Signal Process.*, Las Vegas, NV, 2008, pp. 4669-4672.

[100] Y. Adachi, S. Kawamoto, S. Morishima, and S. Nakamura, "Perceptual similarity measurement of speech by combination of acoustic features," in *Proc. Int. Conf. on Acoust., Speech, Signal Process.*, Las Vegas, NV, 2008, pp. 4861-4864.

[101] K. K. Paliwal, "Interpolation properties of linear prediction parametric representations," in *Proc. EuroSpeech*, Madrid, 1995, pp. 1029-1032.

[102] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Signal Process.*, 1993, SAP-1, (l), pp. 3-14.

[103] Y. Linde, A. Buzo, and R.M. Gray, “An algorithm for vector quantizer design,” *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84-95, 1980.

[104] H. P. Knagenhjelm and W. B. Kleijn, “Spectral dynamics is more important than spectral distortion,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Detroit, MI, 1995, pp. 732-735.

[105] K. S. Lee, “Statistical approach for voice personality transformation,” *IEEE Trans. Audio, Speech and Language Process.*, vol. 15, no. 2, pp. 641-651, 2007.

[106] 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 *Proc. 7th Int. Conf. Spoken Language Process.*, Denver, Col., 2002, pp. 285-288.

[107] Z. Chen and L. H. Zhang, "A ANN based high quality method for voice conversion," in *Proc. Int. Conf. Wireless Commun., Networking and Mobile Computing*, 23-25 Sept., 2010, pp.1-4.

[108] P. Zolfaghari and A. J. Robinson, “A formant vocoder based on mixtures of Gaussians,” in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, Munich, 1997, pp. 1575-1578.

[109] P. Zolfaghari and A. J. Robinson, “Speech coding using a mixture of Gaussians polynomial model,” in *Proc. EuroSpeech*, Budapest, 1999, pp. 1495-1498.

[110] D. Erro, A. Moreno, and A. Bonafonte, “Flexible harmonic/stochastic speech synthesis,” in *Proc. 6th ISCA Speech Synthesis Workshop*, Bonn, 2007.

[111] T. Toda, W. A. Black, and K. Tokuda, “Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, 2005, pp. 9-12.

[112] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Istanbul, 2000, pp. 1315-1318.

[113] T. Najjary, R. Olivier, and T. Chonavel, “A voice conversion method based on joint pitch and spectral envelope transformation,” in *Proc. Int. Conf. Spoken Lang. Process.*, Jeju Island, Korea, 2004, pp. 1225-1228.

[114] L. Zhao and Y. Gao, “Voice conversion adopting SOLAFS,” in *Proc. Eighth ACIS Int. Conf. Software Eng., Artificial Intell. Networking Parallel/Distributed Computing*, Qingdao, China, 2007, pp. 543-548.

[115] H. Ye and S. Young, “Perceptually weighted linear transformation for voice conversion,” in *Proc. EuroSpeech*, 2003, Geneva, Switzerland, pp. 2409-2412.

[116] R. Viswanathan, J. Makhoul, and W. Russell, “Towards perceptually consistent measures of spectral distance,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, 1976, pp. 485-488.

[117] A. Kain and Y. Stylianou, "Stochastic modeling of spectral adjustment for high quality pitch modification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Istanbul, 2000, pp. II949-II952.

[118] H. C. Choi and R. W. King, "On the use of spectral transformation for speaker adaptation in HMM based isolated word speech recognition," *Speech Commun.*, vol. 17, no. 1-2, pp. 131-143, 1995.

[119] A. P. Dempster, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statistical Soc. Series B (Methodological)*, vol. 39, no. 1, pp. 1-38, 1977.

[120] A. Mouchtaris, S. S. Narayanan, and C. Kyriakakis, "Maximum likelihood constrained adaptation for multichannel audio synthesis," in *Proc. 36th Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, 2002, pp. 227-232.

[121] M. M. Wilde and A. B. Martinez, "Probabilistic principal component analysis applied to voice conversion," in *Proc. 38th Asilomar Conf. Signal Sys. Comput.*, Pcific Grove, CA, 2004, pp. 2255-2259.

[122] T. W. Anderson, "Asymptotic theory for canonical correlation analysis,"*J. Multivariate Analysis*, vol. 70, no. 1, pp. 1-29, 1999.

[123] G. M. White and R. B. Neely, "Speech recognition experiments with linear prediction, bandpass filtering, and dynamic programming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 2, pp. 183-188, 1976.

[124] S. Roucos and A. M. Wilgus, "High quality time-scale modification for speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Tampa, FL, 1985, pp. 493-469.

[125] L. Rabiner, R. Schafer, and C. Rader, "The chirp z-transform algorithm," *IEEE Trans. Audio Electroacoust.*, vol. AU-17, pp. 86–92, 1969.

[126] T. Masuda and M. Shozakai, "Cost reduction of training mapping function based on multistep voice conversion," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Honolulu, HI, 2007, pp. 693-696.

[127] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveign, "Restructuring speech representations using a pitch adaptive time-frequency-based F0 extraction possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 34, pp. 187-207, 1999.

[128] Z. H. Ling, Y. P. Wang, Y. Hu, and R. H. Wang, "Modeling glottal effect on the spectral envelop of STRAIGHT using mixture of Gaussians," in *Proc. IEEE Chinese Spoken Language Process.*, Hong Kong, 2004, pp. 73-76.

[129] L. Mesbahi, V. Barreaud, and O. Boeffard, "GMM-based speech transformation systems under data reduction," in *Proc. 6th ISCA Workshop Speech Synthesis*, San Diego, CA, 2007, pp. 119-124.

[130] A. Kumar and A. Verma, "Using phone and diphone based acoustic models for voice conversion: a step toward creating voice fonts," in *Proc. Int. Conf. Multimedia and Expo*, Baltimore, MD, pp. 393-396, 2003.

[131] Z. H. Jian and Z. Yang, "Voice conversion using Viterbi algorithm based on Gaussian mixture model," in *Proc. Int. Symp. Int. Signal Process. and Commun. Syst.*, Xiamen, China, 2007, pp. 32-35.

[132] H. Duxans, A. Bonafonte, A. Kain, and J. van Santen, "Including dynamic and phonetic information in voice conversion systems," in *Proc. Int. Conf. Spoken Lang. Process*, Jeju Island, Korea, 2004, pp. 5-8.

[133] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Munich, 1997, pp. 1611-1614.

[134] O. Turk and L. M. Arslan, "Subband based voice conversion," in *Proc. Int. Conf. Spoken Language Process.*, Denver, Col., 2002, pp. 289-292.

[135] O. Salor, B. C. Pellom, and T. Iloglu, "On developing new text and audio corpora and speech recognition tools for the Turkish language," in *Proc. 7th Int. Conf. Spoken Lang. Process.*, Denver, Colorado, 2002, pp. 349-352.

[136] L. Qin, G. P. Chen, Z. H. Ling, and L. R. Dai, "An improved spectral and prosodic transformation method in STRAIGHT-based voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, 2005, pp. 21-24.

[137] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Munich, 1997, pp. 1303-1306.

[138] A. Verma and A. Kumar, "Voice fonts for individuality representation and transformation," *ACM Trans. Speech and Language Process.*, vol. 2, no. 1, 2005.

[139] J. T. Chien and C. H. Huang, "Bayesian learning of speech duration models," *IEEE Trans. Speech and Audio Process.*, vol. 11, no. 6, pp. 558-567, 2003.

[140] Y. Ueda, M. Hirota, and T. Sakata, "Vowel synthesis based on the spectral morphing and its application to speaker conversion," in *Proc. IEEE Innovative Computing, Inform. and Control*, Beijing, 2006, pp. 738-741.

[141] J. H. Chen and A. Gersho, "Real-time vector APC speech coding at 4800bps with adaptive post filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1987, pp. 2185-2188.

[142] G. Bandoin and Y. Stylianou, "On the transformation of the speech spectrum for voice conversion," in *Proc. Int. Conf. Spoken Language*, Philadelphia, PA, 1996, pp. 1405-1408.

[143] Y. Stylianou, O. Cappe, and E. Moulines, "Statistical methods for voice quality

transformation," in *Proc. EuroSpeech*, Madrid, 1995, pp. 447-450.

[144] A. El Jaroudi and J. Makhoul, "Discrete all-pole modeling for voiced speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,* 1987, pp. 320-323.

[145] T. Galas and X. Rodet, "Generalized functional approximation for source-filter system modeling," in *Proc. EuroSpeech*, Genova, Italy, 1991, pp. 1085-1088.

[146] J. Cohen, T. Kamm, and A. G. Andreou, "Vocal tract normalization in speech recognition compensating for systematic speaker variability," *J. Acoust. Soc. Amer.*, vol. 97, no. 5, pp. 3246-3247, 1995.

[147] M. Mashimo, T. Toda, K. Shikano, and N. Campbell, "Evaluation of cross-language voice conversion based on GMM and straight," in *Proc. EuroSpeech*, Aalborg, Denmark, 2001, pp. 361-364.

[148] Z. Shuang, F. Meng, and Y. Qin, "Voice conversion by combining frequency warping with unit selection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, 2008, pp. 4661-4664.

[149] A. Watanabe, "Formant estimation method using inverse-filter control," *IEEE Trans. Speech and Audio Process.*, vol. 9, no. 4, pp. 317-326, 2001.

[150] J. M. Gutierrez-Arriola, J. M. Montero, J. A. Vallejo, R. Cordoba, R. San-Segundo, and J. M. Pardo, "A new multi-speaker formant synthesizer that applies voice conversion techniques," in *Proc. EuroSpeech*, Aalborg, Denmark, 2001, pp. 357- 360.

[151] D. Rentzos, S. Vaseghi, and Q. Yan, "Voice conversion through transformation of spectral and intonation features," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Montreal, 2004, pp. 21-24.

[152] N. Iwahashi and Y. Sagisaka, "Speech spectrum transformation by speaker interpolation," in *Proc. Int. Conf. on Acoust., Speech, Signal Process.*, Adelaide, SA, 1994, pp. I/461 - I/464.

[153] J. Wouters and M. W. Macon, "A perceptual evaluation of distance measures for concatenative speech synthesis," in *Proc. Int. Conf. Speech Language Process.*, Sydney, 1998.

[154] D. G. Childers, "Glottal source modeling for voice conversion," *Speech Commun.*, vol. 16, no. 2, pp. 127-138, Feb. 1995.

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

[156] H. Duxans and A. Bonafonte, "Residual conversion versus prediction on voice morphing systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Toulouse, France, 2006, pp. I.85-I.86.

[157] K. S. Rao and B. Yegnanarayana, "Voice conversion by prosody and vocal tract modification," in *Proc. 9th Int. Conf. Inform. Technology*, Bhubaneswar, India, 2006,

pp. 111-116.

[158] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. Speech and Audio Process.*, vol. 9, no. 1, pp. 21-29, 2001.

[159] L. M. Arslan and D. Talkin, "Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum," in *Proc. EuroSpeech*, Rhodes, Greece, 1997, pp. 1347 - 1350.

[160] L. Sheng, Y. Juun, and H. Jiancheng, "Voice conversion algorithm using phoneme Gaussian mixture model," in *Proc. Int. Symp. Intelligent Multimedia Video Speech Process.*, Hong Kong, 2004, pp. 5-8.

[161] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, and A. W. Black, "Residual prediction based on unit selection," in *Proc. IEEE Workshop Automat. Speech Recognition and Understanding*, San Juan, Puerto Rico, 2005, pp. 369-674.

[162] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using large speech database," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Atlanta, GA, 1996, vol. 1, pp. 373-376.

[163] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "GMM-based voice conversion applied to emotional speech synthesis," in *Proc. EuroSpeech*, Geneva, 2003, pp. 2401-2404.

[164] L. Lee, C. Tseng, and C. Hsieh, "Improved tone concatenation rules in a formant-based Chinese text-to-speech system," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 3, pp. 287-294, 1993.

[165] C. H. Wu and J. H. Chen, "Automatic generation of synthesis units and prosodic information for Chinese concatenative synthesis," *Speech Commun.*, vol. 35, no. 3-4, pp. 219-237, 2001.

[166] D. H. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Amer.*, vol. 82, no. 3, pp. 737-793, 1987.

[167] K. Silverman, "On customizing prosody in speech synthesis: names and address as a case in point," in *Proc. Workshop Human Language Technology*, 1993, pp. 317-322.

[168] P. Angkititrakul and J. H. L. Hansen, "Advances in phone-based modeling for automatic accent classification," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 2, pp. 634-646, 2006.

[169] Z. Inanoglu and S. Young, "A system for transforming the emotion in speech combining data-driven conversion techniques for prosody and voice quality," in *Proc. EuroSpeech*, Antwerp, Belgium, 2007, pp. 490-493.

[170] M. Schroder, "Emotional speech synthesis: a review," in *Proc. EuroSpeech*, Aalborg, Denmark, 2001, pp. 561-564.

[171] A. Lida, N. Campbell, F. Higuchi, and M. Yasumura, "A corpus-based speech

synthesis system with emotion," *Speech Commun.*, vol. 40, no. 1–2, pp. 161-187, 2003.

[172] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E88-D, no. 3, pp. 502-509, 2005.

[173] E. E. Helander and J. Nurminen, "A novel method for prosody prediction in voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Honolulu, HI, 2007, pp. IV.509-IV.512.

[174] Z. W. Shuang, Z. X. Wang, Z. H. Ling, and R. H. Wang, "A novel voice conversion system based on codebook mapping with phoneme tied weighting," in *Proc. Int. Conf. Spoken Language Process.*, Jeju Island, Korea, 2004, pp. 1197-1200.

[175] B. Gillet and S. King, "Transforming F0 contours," in *Proc. EuroSpeech*, Geneva, 2003, pp. 101-104.

[176] M. Zhang, J. Tao, T. Jilei, and X. Wang, "Text-independent voice conversion based on state mapped codebook," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, Las Vegas, NV, 2008, pp. 4605-4608.

[177] M. M. Hasan, A. M. Nasr, and S. Sultana, "An approach to voice conversion using feature statistical mapping," *Appl. Acoust.*, vol. 66, no. 5, pp. 513-532, 2005.

[178] T. Pfau and G. Ruske, "Estimating the speaking rate by vowel detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Seattle, WA, 1998, pp. 945-948.

[179] E. Moulines and J. Laroche, "Non-parametric techniques for pitch scale and time-scale modification of speech," *Speech Commun.*, vol. 16, no. 2, pp. 175-206, 1995.

[180] W. Hess, *Pitch determination of speech signals*. New York: Springer-Verlag, 1983.

[181] D. J. Mermes, "Pitch analysis," in *Visual Representations of Speech Signals*, M. Cooke, S. Beet, and M. Crawford, Eds. London: Wiley, 1992.

[182] A. Breen, "Speech synthesis models: a review," *Electron. & Commun. Eng. J.*, vol.4, no.1, pp.19-31, 1992.

[183] E. Moulines and Y. Sagisaka (Ed.), "Voice conversion: state of the art and perspectives," Speech Commun., vol. 16, no. 2, pp. 127-138, 1995.

[184] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveignk, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3-4, pp. 187-207, 1999.

[185] K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through combining modified GMM and formant mapping for Mandarin," in *Proc. 2nd Int. Conf. Digital Telecommun.*, San Francisco, CA, 2007, pp. 1-10.

[186] N. Xu, Z. Yang, and H. Guo, "Voice conversion with a strategy for separating

speaker individuality using state-space model," in *Proc. IEEE Int. Conf. Wireless Commun. Networking and Inform. Security*, Beijing, 2010, pp. 298-301.

[187] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, Salt Lake City, UT, 2001, pp. 841-844.

[188] C. Orphanidou, I. M. Moroz, and S. J. Roberts, "Wavelet-based voice morphing," *WSEAS J. Syst.*, vol. 10, no. 3, pp. 3297-3302, 2004.

[189] C. Orphanidou, I. Moroz, and S. J. Roberts, "Multiscale voice morphing using radial basis function analysis," in *Proc. 5th Int. Conf. Algorithms Approximation*, Chester, England, 2005, pp. 61-69.

[190] T. Quatieri and R. McAulay, "Speech transformation based on a sinusoidal representation," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 43, no. 6, pp. 1449-1464, 1986.

[191] F. Huang and J. Yin, "Statistical eigenvoice: speaker features within S+N framework and a way towards language-independent voice conversion," in *Proc. Int. Symp. Intell. Signal Process. Commun.Syst.*, Hong Kong, 2005, pp. 33- 36.

[192] J. Laroche, Y. Stylianou, and E. Moulines, "HNM: A simple, efficient harmonic plus noise model for speech," in *Proc. IEEE Workshop Applicat. Signal Process. Audio and Acoust.*, New Paltz, NY, 1993, pp. 169-172.

[193] Y. Pantazis and Y. Stylianou, "Improving the modeling of the noise part in the harmonic plus noise model of speech," in *Proc. IEEE Int. Conf. Acoust, Speech, Signal Process.*, Las Vegas, NV, 2008, pp. 4609-4612.

[194] Y. Stylianou and A. K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Salt Lake City, UT, 2001, pp. 837–840.

[195] Y. Stylianou, "Removing linear phase mismatches in concatenative speech synthesis," *IEEE Trans. Speech and Audio Process.*, vol. 9, no. 3, pp. 232-239, 2001.

[196] Y. Stylianou, "On the harmonic analysis of speech," in *Proc. IEEE Int. Symp. Circuits Systems*, Monterey, CA, 1998, pp. 5-8.

[197] Y. Stylianou, "A simple and fast way of generating a harmonic signal," *Signal Process. Lett.*, vol. 7, no. 5, pp. 111-113, 2000.

[198] A. Syrdal, Y. Stylianou, L. Garrison, A. Conkie, and J. Schroeter, "TD-PSOLA versus harmonic plus noise model in diphone based speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Seattle, WA, 1998, pp. 273-276.

[199] D. W. Griffin and J. S. Lim, "Multiband-excitation vocoder," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 36, no. 8, pp. 1223-1235, 1988.

[200] S. Seneff, “Real time harmonic pitch detector,” *IEEE Trans. Speech Audio Process.*, vol. 26, no. 4, pp. 358-365, 1978.

[201] T.F. Quatieri and J. McAulay, “Shape invariant time scale and pitch modification of speech,” *IEEE Trans. Signal Process.*, vol. 40, no. 3, pp. 497-510, 1992.

[202] R. Sproat and J. Olive, "An approach to text-to-speech synthesis," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam: Elsevier, 1995, pp. 611–633.

[203] C. Wang, B. Dai, J. Zhang, L. Hui, and L. Yi, “A scheme for high quality linear prediction analysis of speech,” in *Proc. IEEE 3rd Int. Conf. Signal Process.*, Beijing, 1996, pp. 694 – 697.

[204] D. G. Childers and T. H. Hu, “Speech synthesis by glottal excited linear prediction,” *J. Acoust. Soc. Amer.*, vol. 96, no. 4, pp. 2026-2036, 1994.

[205] R. Vergin, D. O’Shaughnessy, and V. Gupta, “Compensated mel frequency cepstrum coefficients,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Atlanta, GA , 1996, pp. 323-326.

[206] O. Cappe, J. Laroche, and E. Moulines, “Regularized estimation of cepstrum envelope from discrete frequency points,” in *Proc. IEEE Workshop Applicat. Signal Process. Audio and Acoust.*, New Paltz, NY, 1995, pp. 213 - 216.

[207] D. O. Shaughnessy, *Speech Communications - Human and Machine*. Hyderabad, India: Universities Press, 2001.

[208] W. Kleijn and K. Paliwal, *Speech Coding and Synthesis.* New York: Marcel Dekker, 1991.

[209] M. Hayes, L. Jae, and A. Oppenheim, “Signal reconstruction from phase or magnitude,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 6, pp. 672-680, 1980.

[210] C. Ma, “Novel criteria of uniqueness for signal reconstruction from phase,” *IEEE Trans. Signal Process.*, vol. 39, no. 4, pp. 989-992, April 1991.

[211] H. Pozidis and A. P. Petropulu, “Signal reconstruction from phase only information and application to blind system estimation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Munich, 1997, pp. 1869-1872.

[212] T. F. Quatieri and A. Oppenheim, “Iterative techniques for minimum phase signal reconstruction from phase or magnitude,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 6, pp. 1187-1193, 1981.

[213] R. B. Blahut, *Fast Algorithms for Digital Signal Processing*. Boston, MA: Addision-Wesley, 1985.

[214] O. J. Smith, *Introduction to Digital Filters with Audio Applications*. W3K Publishing, 2007.

[215] Y. Stylianou, J. Laroche, and E. Moulines, “High-quality speech modification based on a harmonic + noise model,” in *Proc. EuroSpeech*, Madrid, Spain, 1995, pp. 451-454.

[216] E. Boucheron and P. L. De Leon, "On the inversion of mel-frequency cepstral coefficients for speech enhancement applications,” in *Proc. Int. Conf. Signals and Electronic Systems (ICSES)*, Krakow, Poland, 2008, pp. 485-488.

[217] J. W. Picone, “Signal modeling techniques in speech recognition,” *Proc. IEEE*, vol. 81, no. 9, 1993, pp. 1215-1247.

[218] E. Moulines and F. Charpentier, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” *Speech Comm.*, vol. 9, no. 5-6, pp. 453–467, 1990.

[219] M. Vasilakis and Y. Stylianou, “A mathematical model for accurate measurement of jitter,” in *Proc. Fifth Int. Workshop on Models and Analysis of Vocal Emissions for Biomedical Applicat.*, Florence, Italy, 2007, pp. 7-10.

[220] M. Farrus, J. Hernando, and P. Ejarque, “Jitter and shimmer measurements for speaker recognition,” in *Proc. InterSpeech*, Antwerp, Belgium, 2007, pp. 778-781.

[221] R. E. Slyh, W. T. Nelson, and E. G. Hansen, "Analysis of mrate, shimmer, jitter, and F_0 contour features across stress and speaking style in the SUSAS database,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Phoenix, Arizona, 1999, vol. 4, pp. 2091-2094.

[222] *F-J Electronics A/S, Electrograph (type EG 810)*, Vangede Bygade 114, DK-2820 Gentofte, Denmark.

[223] R. H. Colton and E. G. Conture, “Problems and pitfalls of electro-glottography,” *Journal of Voice*, vol. 4, no. 1, pp. 10-24, 1990.

[224] L. Meshabi, V. Barreaud, and O. Boeffard, “Comparing GMM-based speech transformation systems”, in *Proc. InterSpeech*, Antwerp, Belgium, 2007, pp. 1989-1992.

[225] G. M. Philips, *Interpolation and Approximation by Polynomials*. New York: Springer-Verlag, 2003.

[226] R. L. Branham Jr., *Scientific Data Analysis: An Introduction to Overdetermined Systems*. New York: Springer-Verlag, 1990.

[227] V. Pratt, “Direct least-squares fitting of algebraic surfaces,” *Computer Graphics*, vol. 21, no. 4, pp. 145–152, July 1987.

[228] R. Bellman and R. Kalaba, “On adaptive control processes,” *IRE Trans. Automatic Control*, vol. 4, no. 2, pp. 1-9, 1959.

[229] S. Salvador and P. Chan, “Toward accurate dynamic time warping in linear time and space,” *Intell. Data Anal.*, vol. 11, no. 5, pp. 561-580, 2007.

[230] D. J. C. MacKay, “An example inference task”, in *Information Theory, Inference and Learning Algorithms*. Cambridge, UK: Cambridge University Press, 2003.

[231] ITU, “Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” *ITU-T Rec*., P.862, 2001.

[232] P. K. Lehana, “Voice conversion using multivariate polynomial modeling: demo,” 2013, [online] Available: www.ee.iitb.ac.in/~spilab/material/parveen_ lehana/voice_conversion_demo.

[233] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, “Voice conversion with smoothed GMM and MAP adaptation,” in *Proc. EuroSpeech*, Geneva, Switzerland, 2003, pp. 2413-2416.

[234] Y. Kang, Z. Shuang, T. Jianhua, W. Zhang, and B. Xu, “A hybrid GMM and codebook mapping method for spectral conversion,” in *Proc. of Affective Computing Intelligent Interaction*, 2005, pp. 303-310.

[235] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Process.*, vol. 10, no.1–3, pp. 19–41, 2000.

[236] Y. L. Tong, *The Multivariate Normal Distribution*. New York: Springer-Verlag, 1990.

[237] M. I. Schlesinger and V. Hlavac, *Ten Lectures on Statistical and Structural Pattern Recognition*. Computational Imaging and Vision. Kluwer Academic Publishers, 2002.

[238] R. A. Redner and H. F. Walker, “Mixture densities, maximum likelihood and the EM algorithm,” *SIAM Rev*., vol. 26, pp. 195–239, 1984.

[239] M. A. T. Figueiredo and A. K. Jain, “Unsupervised learning of finite mixture models,” *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 24, no.3, pp. 381-396, 2002.

[240] M. N. Stuttle, “A Gaussian mixture model spectral representation for speech recognition,” Ph.D. Thesis, Hughes Hall and Cambridge University Engineering Department, 2003.

[241] H. Duxans, “A Gaussian mixture model spectral representation for speech recognition,” Ph.D. Thesis, Technical University of Catalonia, Spain, 2006.

[242] T. H. Falk and W. Y. Chan, “Objective speech quality assessment using Gaussian mixture models,” in *Proc. 32nd Biennial Symposium on Commun.*, Kingston, Ont., Canada, 2004, pp. 169-171.

[243] S. Wang, A. Sekey, and A. Gersho, “An objective measure for predicting subjective quality of speech coders,” *IEEE J. Sel. Areas Commun.*, vol. 10, no. 5, pp. 819-829, 1992.

[244] J. G. Beerends and J. A. Stemerdink, “A perceptual speech-quality measure based on

a psychoacoustic sound representation," *J. Audio Eng. Soc.*, vol. 42, no. 3, pp. 115 - 123, 1994.

[245] S. Voran, "Objective estimation of perceived speech quality," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 4, pp. 371-382, 1999.

[246] A. W. Rix and M. P. Hollier, "The perceptual analysis measurement system for robust end-to-end speech quality assessment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Istanbul, Turkey, 2000, pp. 1515-1518.

[247] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Salt Lake City, 2001, UT, pp. 749-752.

[248] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual evaluation of speech quality, the new ITU standard for end-to-end speech quality assessment part II Psychoacoustic model," *J. Audio Eng. Soc.*, vol. 50, no. 10, pp. 765-778, 2002.

[249] L. Ding and R. A. Goubran, "Assessment of effects of packet loss on speech quality in VoIP," in *Proc. IEEE Int. Workshop Haptic, Audio and Visual Environments and their Appl.*, Ottawa, 2003, pp. 49-54.

[250] T. Watanabe, H. Nagabuchi, and N. Kitawaki, "A word-selecting method for intelligibility assessment of synthesized speech by rule," *Trans. IEICE Japan*, vol. 71A, no. 3, pp. 616-623, 1988.

[251] J. P. Eagan, "Articulation testing methods," *Laryngoscope*, vol. 58, no. 9, pp. 955-991, 1948.

[252] G. Fairfanks, "Test of phonemic differentiation the rhyme test," *J. Acoust. Soc. Amer.*, vol. 30, no. 7, pp. 596-600, 1958.

[253] ITU, "Methods for subjective determination of transmission quality," *Tech. Rep. ITU-T Recommendation P.800, ITU,* Aug. 1996.

[254] T. Ganchev, A. Lazaridis, I. Mporas, and N. Fakotakis, *Performance Evaluation for Voice Conversion Systems.* Berlin: Springer, 2008.

[255] N. Kitawaki and H. Nagabuchi, "Quality assessment of speech coding and speech synthesis systems," *IEEE Commun. Mag.*, vol. 26, no. 10, pp. 36-44, 1988.

[256] CCITT, "Telephone transmission quality" *Red Book, ITU,* Geneva, vol. 5, VIIIth Plenary Assembly, 1985.

[257] D. J. Goodman and R. Nash, "Subjective quality of the same speech transmission conditions in seven different countries," *IEEE Trans. Commun.*, vol. 30, no. 4, pp. 642-654, 1982.

[258] P. Combescure, A. Le Guyader, and A. Gilloire, "Quality evaluation of 32 kbit/s coded speech by means of degradation category ratings," in *Proc. IEEE Int. Conf.*

Acoust., and Speech, Signal Process., Paris, 1982, pp. 988-991.

[259] J. Kreiman and G. Papcun, "Comparing, discrimination and recognition of unfamiliar voices," *Speech Commun.*, vol. 10, no. 3, pp. 265-275, 1991.

[260] S. Meister and R. Wiggins, "Quality comparison measure for linear predictive systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, 1976, pp. 107-109.

[261] J. Makhoul, R. Viswanathan, and W. Russell, "A framework for the objective evaluation of vocoder speech quality," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, 1976, pp. 103-106.

[262] T. P. Barnwell, "Correlation analysis of subjective and objective measures for speech quality," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Denver, Col., 1980, pp. 706-709.

[263] T. P. Barnwell and W. D. Voiers, "An analysis of objective measures for user acceptance of voice communications systems," *Georgia Inst. of Tech. Atlanta School of Elect. Eng., Final Rep.*, no. ADA089210, 1979.

[264] R. Viswanathan and J. Makhoul, "Quantization properties of transmission parameters in linear predictive systems," *IEEE Trans. Audio Speech Signal Process.*, pp. 309-321, 1975.

[265] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low bit-rate speech coding systems," *IEEE J. Sel. Areas Commun.*, vol. 6, no. 2, pp. 262-273, 1988.

[266] Y. Hu and Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229-238, 2008.

[267] P. C. Mahalanobis, "On the generalised distance in statistics," in *Proc. National Institute of Sciences of India*, 1936, pp. 49–55.

[268] T. Takeshita, S. Nozawa, and F. Kimura, "On the bias of Mahalanobis distance due to limited sample size effect," in *Proc. 2nd IEEE Int. Conf. Document Analysis and Recognition*, Tsukuba Science City, Japan, 1993, pp. 171-174.

[269] J. C. T. B. Moraes, M. O. Seixas, F. N. Vilani, and E. V. Costa, "A real time QRS complex classification method using Mahalanobis distance," in *Proc. Computers in Cardiology*, Memphis, Tenn., 2002, pp. 201-204.

[270] T. Kamei, "Face retrieval by an adaptive Mahalanobis distance using a confidence factor," in *Proc. IEEE Int. Conf. Image Process.*, 2002, pp. 153-156.

[271] G. Chen, H. G. Zhang, and J. Guo, "Efficient computation of Mahalanobis distance in financial hand-written Chinese character recognition," in *Proc. IEE Int. conf. Machine Learning and Cybernetics*, Hong Kong, 2007, vol. 4, pp. 2198-2201.

[272] J. M. Yih, D. B. Wu, and C. C. Chen, “Fuzzy C-mean algorithm based on Mahalanobis distance and new separable criterion,” in *Proc. IEEE Int. Conf. Machine Learning and Cybernetics*, Hong Kong, 2007, pp. 1851-1855.

[273] J. P. Campbell, “Speaker recognition: a tutorial,” in *Proc. IEEE*, vol. 85, pp. 1437-1462, 1997.

[274] R. Viswanathan, W. Russel, and J. Makhoul, “Objective speech quality evaluation of narrowband LPC vocoders,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Tulsa, Okl., 1978, pp. 591- 594.

[275] B. Pellom and J. Hansen, “Spectral normalization employing hidden Markov modeling of line spectrum pair frequencies,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Munich, 1997, pp. 943-946.

[276] N. Ma, M. Bouchard, and R. A. Goubran, "A perceptual Kalman filtering-based approach for speech enhancement," *in Proc. Int. Symp. Signal Process. Applicat.*, 2003, vol. 1, pp. 373- 376.

[blank]

ACKNOWLEDGEMENTS

With deep regards and profound respect, I take this opportunity to express my deep sense of gratitude and indebtedness to my supervisor Prof. P. C. Pandey for unconditional support at both personal and academic levels, invaluable guidance, motivation, and constructive criticism which have made this work possible. I am thankful to Prof. V. M. Gadre and Prof. P. Rao, members of the research progress committee for my thesis, for their valuable suggestions and encouragements at various stages of the work.

I would like to thank everyone in the Signal Processing & Instrumentation (SPI) Lab for providing a cordial environment. My special thanks go to my friends, Alice, Dakshayani, Milind, Vinod, Pandurang, Jayan, Arup, Santosh, Mohan, Nataraj, Jagbandhu, Rajath, Sudipan, Nitya, Dinesh, Srikant, and Vidyadhar for encouragement and support, whenever I needed it. It indeed has been an honor and privilege to work and live in the wonderful environment of the Institute. I have travelled a long journey at IIT Bombay, and I may have missed mentioning some of my good friends and I apologize to them.

I am really indebted to my parents, parents-in law, my wife Santoresh, and other family members for being so supportive and patient, especially for boosting my psychological strength to reach the target.

I dedicate this thesis to my mother, may be she is the only one busy in counting each and every moment I have been spending at IIT Bombay for getting the PhD degree. I think it is her belief and the inspiration derived from the following lines written by Robert Frost that I could cover this long journey to complete the thesis.

"*The woods are lovely, dark, and deep,*
But I have promises to keep,
And miles to go before I sleep,
And miles to go before I sleep."

Parveen Kumar Lehana

[blank]

AUTHOR'S RESUME

Parveen Kumar Lehana received the M.Sc. degree in Electronics from Kurukshetra University, Kurukshetra in 1992. After teaching Electronics in a degree college for two years and qualifying JRF-NET, he joined as lecturer in P. G. Department of Physics and Electronics, University of Jammu, Jammu. Presently he is Associate Professor in the same department and pursuing Ph.D. in Electrical Engineering at the Indian Institute of Technology Bombay, India. His research interests include digital signal processing, speech synthesis, and voice conversion.

LIST OF PUBLICATIONS

Journals (Abstracts)

1. P. K. Lehana and P. C. Pandey, "Speech synthesis with pitch modification using harmonic plus noise model," *J. Acoust. Soc. Am*., vol. 114, no. 4, pp. 2395, 2003.
2. P. K. Lehana and P. C. Pandey, "Speech enhancement during analysis-synthesis by harmonic plus noise model," *J. Acoust. Soc. Am*., vol. 120, pp. 3039, 2006.

International Conferences

3. P. K. Lehana and P. C. Pandey, "The effect of SNR and GCI's purturbation on speech synthesis with harmonic plus noise model," in *Proc. 7th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2003)*, Orlando, USA, 27-30 July, 2003.
4. P. K. Lehana and P. C. Pandey, "Effect of GCI perturbation on speech quality in Indian languages," in *Proc. IEEE TENCON-2003*, 15-17 October, 2003, Bangalore, India, pp. 959-963.
5. P. K. Lehana and P. C. Pandey, "Harmonic plus noise model based speech synthesis in Hindi and pitch modification," in *Proc. 18th Int. Cong. Acoust.*, 4-9 April, 2004, Kyoto, Japan, pp. 3333-3336.

National Conferences

6. P. K. Lehana and P. C. Pandey, "A low cost impedance glottograph and glottal pitch analyzer," in *Proc. ICBME*, IISc, Bangalore, India, 2001.

7. P. K. Lehana and P. C. Pandey, “Speech synthesis in Indian languages,” in *Proc. Int. Conf. Universal Knowledge and Language - 2002*, Goa, India, 25-29 Nov., 2002, paper no. 14.
8. P. K. Lehana and P. C. Pandey, “Improving speech synthesis in Indian languages,” in *Proc. Workshop Spoken Language Process.*, 9-11 January, 2003, TIFR, Mumbai, India, pp. 149-155.
9. P. K. Lehana and P. C. Pandey, “Perturbation in GCIs and speech quality for pitch synchronous synthesis,” in *Proc. Symp. Frontiers of Research on Speech and Music - 2003*, IIT Kanpur, Kanpur, India, February 15-16, 2003, pp. 86-94.
10. P. K. Lehana, P. C. Pandey, and R. Gupta, “Use of harmonic plus noise model for reduction of self leakage in electroalaryngeal speech,” in *Proc. Int. Conf. Systemics, Cybernetics and Informatics (SCI 2004)*, 12-15 February, 2004, Hyderabad, India, pp. 366-370.
11. P. K. Lehana and P. C. Pandey, “Speaker transformation using quadratic surface interpolation,” in *Proc. 14th National Conf. Commun.*, 1-3 Feb., 2008, IIT Bombay, India, pp. 190:194.
12. P. K. Lehana and P. C. Pandey, “Transformation of short-term spectral envelope of speech signal using multivariate polynomial modeling,” in *Proc. 17th National Conf. Commun.*, 1-3 Feb., 2011, IISc, Bangalore, India, paper SpPrII.2.

Indian Institute of Technology Bombay

CERTIFICATE OF COURSE WORK

This is to certify that Mr. Parveen Kumar Lehana (Roll No. 00407304) was admitted to the candidacy of Ph.D. Degree in July 2000, after successfully completing all the courses required for the Ph.D. Degree Programme. The details of the course work done are given below.

Sr. No.	Course No.	Course Name	Credits
1	EE 600	Mini Project	10
2	EE 603	Digital Signal Processing, and its Applications	6
3	EE 610	Image Processing	6
4	BM 632	Medical Instrumentation	6
5	EE 679	Speech Processing	6
6	EE 712	Embedded Systems Design	6
7	EE 801	Seminar	4
8	EE 802	Seminar	4
		Total Credits	48

I.I.T. Bombay
Dated:

Dy. Register (Academic)