

**SPEECH CONVERSION AND ITS APPLICATION TO
ALARYNGEAL SPEECH ENHANCEMENT**

by
Ning Bi

A Dissertation Submitted to the Faculty of the
DEPARTMENT OF SPEECH AND HEARING SCIENCES
In Partial Fulfillment of the Requirements
For the Degree of
DOCTOR OF PHILOSOPHY
In the Graduate College
THE UNIVERSITY OF ARIZONA
1 9 9 5

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700 800/521-0600

**SPEECH CONVERSION AND ITS APPLICATION TO
ALARYNGEAL SPEECH ENHANCEMENT**

by
Ning Bi

A Dissertation Submitted to the Faculty of the
DEPARTMENT OF SPEECH AND HEARING SCIENCES
In Partial Fulfillment of the Requirements
For the Degree of
DOCTOR OF PHILOSOPHY
In the Graduate College
THE UNIVERSITY OF ARIZONA
1 9 9 5

UMI Number: 9604516

**UMI Microform 9604516
Copyright 1995, by UMI Company. All rights reserved.**

**This microform edition is protected against unauthorized
copying under Title 17, United States Code.**

UMI

**300 North Zeeb Road
Ann Arbor, MI 48103**

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Final Examination Committee, we certify that we have read the dissertation prepared by Ning Bi

entitled Speech Conversion and Its Application to Alaryngeal

Speech Enhancement

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

Yingyong Qi
Yingyong Qi

May 9, 1995
Date

Theodore J. Glattke
Theodore J. Glattke

May 9, 1995
Date

Thomas Shipp
Thomas Shipp

May 9, 1995
Date

Date

Date

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copy of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Yingyong Qi
Dissertation Director
Yingyong Qi

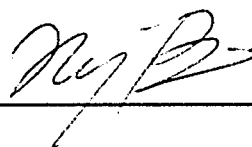
June 9, 1995
Date

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: _____

A handwritten signature in cursive script, appearing to read 'M. J. B.', is written over a horizontal line.

ACKNOWLEDGMENTS

I am deeply grateful to my advisor, Dr. Yingyong Qi, for his professional guidance, constant support, friendly encouragement, and his devotion of much time and energy in the undertaking of this research. Grateful appreciation is also extended to Dr. Theodore J. Glatke, Dr. Thomas J. Hixon, Dr. Bobby R. Hunt, Dr. Thomas Shipp, and Dr. Robin N. Strickland for their input and support as members of my Graduate Committee. I acknowledge all those who volunteered to participate as speakers and listeners in this study, and I thank the professors, colleagues, and Department staffs for their assistance in this work. My deep appreciation goes to my wife Jianhui Shi and my parents for their all-out support. Finally, I cherish the hope and believe that my grandmother will know of my accomplishment.

TABLE OF CONTENTS

LIST OF FIGURES	7
LIST OF TABLES	9
1 INTRODUCTION	11
1.1 Background	12
1.2 Objectives	15
1.3 Summary of Contributions	16
1.4 Overview	17
2 SPEECH ANALYSIS-SYNTHESIS	18
2.1 Source-Filter Model of Speech Production	19
2.2 Linear Predictive Analysis-Synthesis	25
2.3 Homomorphic Speech Processing	30
2.4 Vowel Detection and Pitch Period Estimation	38
2.5 Glottal Flow Model and Its Approximation	38
2.6 Dynamic Time Warping in Pattern Alignment	43
2.7 Voicing Source Replacement and Synthesis	47
3 SPEECH CONVERSION	52
3.1 VQ-Based Spectral Conversion	54
3.1.1 Vector Quantization and Fuzzy Vector Quantization	55
3.1.2 Feature Space Mapping and Fuzzy Feature Space Mapping	60
3.2 LMR-Based Spectral Conversion	70
3.2.1 Linear Multivariate Regression	70
3.2.2 Capacity of Multi-Subset LMR	72
4 MODIFICATIONS OF SPEECH CONVERSION METHODS	78
4.1 Modification of VQ-Based Conversion Method	78
4.1.1 Formant Enhancement Using Chirp Z-Transform	79
4.1.2 Formants Enhancement Using Cepstral Weighting	82
4.2 Modification of LMR-Based Method	86
4.2.1 Overlapped Multi-Subset Training Approach	86
5 APPLICATION TO ALARYNGEAL SPEECH ENHANCEMENT	91
5.1 Subjects and Recordings	91

5.2	System Implementation	95
5.2.1	Speech Analysis	95
5.2.2	Voicing Source Replacement	95
5.2.3	Spectral Conversion	96
5.2.3.1	VQ-Based Conversion System	96
5.2.3.2	LMR-Based Conversion System	101
5.3	Perceptual Evaluation	104
5.4	Evaluation Results	105
5.4.1	Listener Reliability	105
5.4.2	Summary of Evaluation Scores	105
6	SUMMARY AND CONCLUSION	108
	REFERENCES	111

LIST OF FIGURES

2.1	Source-filter model of speech production	20
2.2	Temporal signal of source-filter model	22
2.3	Spectrum of source-filter model	23
2.4	Linear prediction model	26
2.5	Homomorphic processing	32
2.6	The use of homomorphic filtering to extract the cepstrum and the impulse response of the system	35
2.7	Spectra of LPC coefficients, cepstral coefficients, and LPC cepstral coefficients	37
2.8	LF model of glottal flow	40
2.9	Approximation of LF model	44
2.10	A schematic illustration of dynamic programming	46
2.11	Block diagram of the LPC analysis-synthesis system to improve ala- ryngeal speech	49
2.12	Block diagram of the homomorphic processing to improve alaryngeal speech	50
2.13	Block diagram of the combination of LPC and homomorphic process- ing to improve alaryngeal speech	51
3.1	Vector quantization and fuzzy vector quantization	57
3.2	Distortion of vector quantization and fuzzy vector quantization	58
3.3	Illustration of VQ and FVQ processing by three dimensional spectra	59
3.4	VQ-based feature space mapping	63
3.5	VQ-based feature space fuzzy mapping with and without sufficient training	65
3.6	Distortion of fuzzy mapping	66
3.7	Distribution of the number of the training vector-pairs in each codeword	68
3.8	Illustration of VQ-based speech conversion by three dimensional spectra	69
3.9	LMR-based feature space mapping	73
3.10	Distortion of multi-subset LMR mapping	75
3.11	Multi-subset LMR with and without sufficient training	77
4.1	Example of formant enhancement using the chirp z-transform	81
4.2	Rectangular and sine windows and their corresponding spectra	83
4.3	Example of formant enhancement using cepstral weighting	84
4.4	Illustration of formant enhancement by three dimensional spectra . . .	85

4.5	Illustration of overlapped multi-subset training approach for LMR-based method	87
4.6	LMR-based feature space mapping with overlapped subset training approach	89
4.7	Illustration of LMR-based spectral conversion with and without overlapped training approach by three dimensional spectra	90
5.1	Formants of the male alaryngeal subject and the normal subject . . .	93
5.2	Vowel formant space of the alaryngeal subject and the normal subject	94
5.3	Period determination by cepstral coefficients	97
5.4	Illustration of DTW processing by three dimensional spectra	99
5.5	Block diagram of learning phase in the VQ-based spectral conversion approach	100
5.6	Block diagram of conversion-synthesis phase in the VQ-based spectral conversion approach	101
5.7	Block diagram of learning phase in the LMR-based spectral conversion approach	103
5.8	Block diagram of conversion-synthesis phase in the LMR-based spectral conversion approach	103

LIST OF TABLES

5.1	Number and percentage of responses preferring conditions of word in the first column (subject 1)	106
5.2	Number and percentage of responses preferring conditions of word in the first column (subject 2)	107

ABSTRACT

In this investigation, a vector quantization (VQ)-based speech conversion algorithm and a linear multivariate regression (LMR)-based speech conversion algorithm were modified, and the modified algorithms were applied to the enhancement of alaryngeal speech. The modifications were aimed at reducing the spectral distortion (bandwidth increase) in the VQ-based system and the spectral discontinuity in the LMR-based system. The spectral distortion in the VQ-based algorithm was compensated by formant enhancement using chirp z-transform and cepstral weighting. The spectral discontinuity in the LMR-based system was minimized by the use of overlapped subsets during the constructing of conversion mapping function. These modified algorithms were evaluated using simulated data and speech samples. Results of the evaluations indicated that the modified algorithms reduced conversion distortions. These modified algorithms were also used for the enhancement of alaryngeal speech. Results of perceptual evaluation indicated that listeners generally preferred to listen to the enhanced speech samples.

CHAPTER 1

INTRODUCTION

The normal voicing source is a stream of modulated airflow. This airflow, supplied by the respiratory system and modulated by the actions of the larynx, is essential for producing speech. Unfortunately, laryngeal cancer may necessitate a partial or total removal of the larynx, resulting in a fundamental change of the speech production mechanism.

Two widely used methods of restoring speech following total laryngectomy are esophageal and tracheoesophageal speech. Both esophageal and tracheoesophageal speech rely upon a common voicing source: a surgically altered and reconstructed pharyngo-esophageal (PE) segment. These two forms of alaryngeal speech rely upon entirely different sources of air supply and airway status. For example, esophageal speech is produced with an open trachea stoma and an open lower airway. The air supply available to sustain esophageal voicing is limited to relatively small volumes of air captured in the mouth and pharynx and delivered into the esophagus under positive pressure. Tracheoesophageal speech is produced with an occluded trachea stoma and a closed lower airway. The air supply available to sustain tracheoesophageal voicing is the large volume of pulmonary air that is shunted into the esophagus via the tracheoesophageal puncture.

A number of studies have addressed the physical properties of esophageal and tracheoesophageal speech (Weinberg and Bennett, 1972; Sisty and Weinberg, 1972; Bennett and Weinberg, 1973; Smith et al., 1978; Weinberg et al., 1980; Weinberg, 1982; Weinberg, 1986; Robbins et al., 1984; Nord and Hammarberg, 1989; Trudeau and Qi, 1990; Qi and Weinberg, 1991; Qi and Weinberg, 1995). Both esophageal and tracheoesophageal types of alaryngeal speech are characterized by low average fundamental frequency and large perturbations in fundamental frequency (Robbins et al., 1984; Trudeau and Qi, 1990). Some other properties of alaryngeal speech such as formant frequencies and spectral slope may also differ significantly from normal speech (Sisty and Weinberg, 1972; Qi and Weinberg, 1991). Perceptually, alaryngeal speech is often described as rough-hoarse and strain-tense. The overall objective of the present study is to develop a spectral-conversion-based method that will enhance the quality of alaryngeal speech.

1.1 Background

To enhance the quality of alaryngeal speech, Qi (Qi, 1990) attempted replacing the voicing source of alaryngeal speech using linear predictive coding (LPC) techniques. It was demonstrated that a LPC analysis-synthesis method could be used to separate vocal tract transfer functions from source functions of vowels produced by alaryngeal individuals. Vowels synthesized with reconstructed vocal tract transfer functions and totally synthetic voicing excitations improved source-related properties over those

present in the original vowels. A recent extension of this work aimed at developing a system to enhance esophageal and tracheoesophageal speech demonstrated that words spoken by female esophageal and tracheoesophageal talkers could also be enhanced by means of LPC-based analysis and synthesis methods (Qi et al., 1995). There are two basic assumptions under these studies: articulatory-based acoustic features of alaryngeal speech are not significantly modified by laryngectomy; and vocal tract transfer functions of alaryngeal speech could be accurately determined using LPC analysis.

These assumptions should be applicable to most alaryngeal speech because only the larynx is surgically removed during laryngectomy. In some special cases, however, these assumptions may not be valid. For example, the formant frequencies of alaryngeal speech may be significantly shifted upward due to the possible surgical shortening of the vocal tract (Sisty and Weinberg, 1972). Larynx removal may also alter other articulatory behaviors because of the disrupted muscular support for the tongue (Weinberg, 1986). In these cases, both source- and articulation-related parameters of alaryngeal speech need to be modified to achieve enhancement.

It has been documented that spectral conversion is a feasible technique for modifying articulation-related parameters of speech (Shikano et al., 1986; Abe et al., 1988; Nakamura and Shikano, 1989; Abe et al., 1990; Abe, 1991; Shikano et al., 1991; Valbret et al., 1992). Spectral conversion was originally used for speaker adaptation in speech recognition systems, where the spectral information is described by

a codebook. The speech of a given speaker was adapted to the system by mapping the spectral codebook of the speaker to a reference codebook prior to recognition. This adaptation was reported to be effective in minimizing the effect of inter-speaker variations during speech recognition.

The technique of spectral conversion was also used in normal voice conversion systems (Abe et al., 1988; Abe et al., 1990; Abe, 1991). To accomplish the voice conversion, the spectral spaces of an input speaker and a target speaker were reduced to, and represented by two codebooks that were obtained using vector quantization (VQ) techniques. The mapping rules between the two codebooks were generated using a supervised learning procedure. Voice transformation was accomplished by applying these mapping rules in a LPC-based analysis and synthesis system.

This VQ-based spectral conversion method has two major sources of error/distortion. First, the reduction of a continuous spectral space into a discrete codebook inevitably introduces quantization noise, which is the difference between a given spectrum and its corresponding codeword (representative spectrum) in the codebook. Second, the codeword typically is an average of a small set of spectra and, thus, tends to have a larger bandwidth than the original spectrum. In an effort to reduce quantization noise, Shikano et. al. (1991) proposed a fuzzy vector quantization (FVQ) method, in which an input spectrum was coded as a weighted interpolation of a set of codewords. This weighted interpolation undoubtedly has the potential to reduce quantization noise because the spectral space is now approximated by many inter-connected

lines between codewords, rather than by a point grid of codewords. The weighted interpolation, however, tends to increase further the bandwidth of the final coded spectrum.

Valbret et. al. used a linear multivariate regression (LMR) approach for spectral conversion (Valbret et al., 1992). In this approach, the spectral space was partitioned by a few large clusters and the spectrum within each cluster was mapped continuously. The mapping matrix was obtained using procedures of least-square approximation. Because the mapping in any given region of the spectral space is continuous, the distortion due to quantization noise and spectral averaging were essentially eliminated. The transitions between clusters, however, can be discontinuous resulting in audible clicks in the converted speech output.

Despite of the problems of spectral averaging in VQ-based system and transition discontinuity in LMR-based system, it has been reported that the conversions were successful in that the converted speech is perceptually more close to the target than to the original speech. Speech quality was not a major concern in these studies. However, the quality of speech would be the primary concern when using spectral conversion for speech enhancement.

1.2 Objectives

The goal of this investigation is to develop speech conversion algorithms for the enhancement of alaryngeal speech. The specific objectives are:

- To modify the VQ-based method to reduce conversion distortions due to formant bandwidth increase.
- To modify the LMR-based method to reduce transition discontinuities during speech conversion.
- To evaluate and compare the performance of VQ- and LMR-based systems.
- To determine if these modified spectral conversion methods can be used for alaryngeal speech enhancement.

1.3 Summary of Contributions

The main contributions of this work include:

- An algorithm has been developed to enhance the formant frequencies in the VQ-based speech conversion system.
- An algorithm has been developed to smooth the spectral transitions in the LMR-based speech conversion system.
- It has been demonstrated that these modified speech conversion systems could be used for the enhancement of alaryngeal speech.

These contributions provide method and guidance to the future development of speech enhancement systems for people with articulatory deficits.

1.4 Overview

Theories and mathematical procedures of speech analysis-synthesis are reviewed and described in chapter 2. Three methods of constructing vocal tract transfer function from speech signals are introduced.

Chapter 3 introduces the VQ- and LMR-based speech conversion algorithms. Properties of vector quantization, fuzzy vector quantization, fuzzy mapping and linear multivariate regression algorithms are studied using simulated data. Problems of these algorithms are identified and schemes for improvement are proposed.

Chapter 4 presents formants enhancement and the overlapped multi-subset training approaches developed in this work for improving the speech conversion systems.

Chapter 5 describes the implementations of the VQ-based and the LMR-based systems. Procedures and results of perceptual evaluations are presented.

Chapter 6 contains summary and conclusion of this work.

CHAPTER 2

SPEECH ANALYSIS-SYNTHESIS

In the present study, the enhancement of alaryngeal speech involves the following general steps:

1. decomposing alaryngeal speech into excitations and spectral transfer functions using linear predictive analysis;
2. replacing the excitations of alaryngeal speech with synthetic excitations;
3. replacing the transfer functions of alaryngeal speech with normal transfer functions;
4. synthesizing speech using the replaced excitations and transfer functions.

Our strategy is based on the source-filter model of speech production, which allows prosodic and spectral modifications to be performed independently. The methods used to decompose and recombine the speech signal are the linear predictive analysis-synthesis and homomorphic processing. Relevant theories and procedures of speech analysis-synthesis are reviewed and described in this chapter. Attention is focused on the concepts, definitions, and algorithms that are applied to this work. Finally, three methods of constructing vocal tract transfer function from speech signals, which can

be used to enhance alaryngeal speech by replacing alaryngeal voicing sources, are introduced.

2.1 Source-Filter Model of Speech Production

The mechanism of normal speech production has been well studied (Stevens and House, 1955; Fant, 1960; Fant, 1981; Flanagan, 1972; Ishizaka and Flanagan, 1972; Stevens, 1989). A voiced source is an airflow interrupted by the vibration of the vocal folds. The vibration can be well controlled by the actions of the larynx, which is essential for producing natural sounding voices. The voice source signal consists of a quasi-periodic pulse train of air characterized by rich harmonies with a $12dB/octave$ spectral decaying. This signal is then modulated in components of harmonies while passing through the vocal tract. An unvoiced source is generated by air turbulence resulting from airflow passing through a narrow constriction in the vocal tract.

The vocal tract begins at the vocal folds or glottis and ends at the lips; it consists of the pharynx and the oral cavity. The vocal tract is a non-uniform acoustic tube that changes in cross sectional area from the glottis to the lips and also varies in shape as a function of time. The positions of the tongue, jaw, lips, and velum determine the acoustic properties of the tube. The vocal tract acts as an acoustic resonant cavity that enhances some of the harmonies in the voice source signal and attenuates others. The nasal tract is acoustically coupled to the vocal tract to produce the nasal sounds

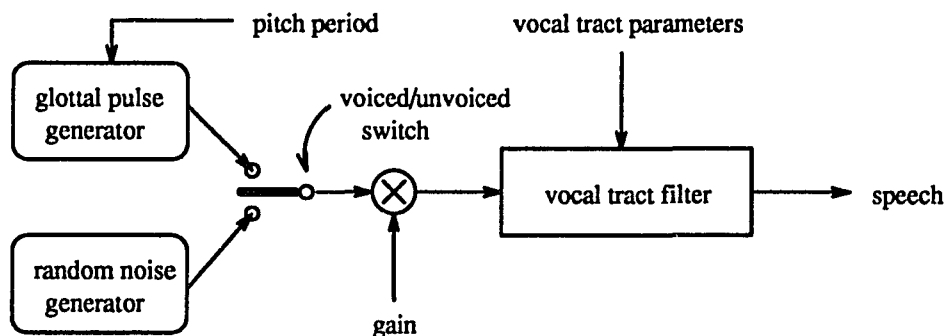


Figure 2.1: Source-filter model of speech production

of speech. The effects of the acoustic radiation from the mouth can be modeled as a high-pass filter with a $6\text{dB}/\text{octave}$ spectral gain in sound pressure level.

Acoustic theory of normal speech production treats the speech signal as an output of a linear source-filter system in which the vocal tract acts as the filter and the airflow modulated by the vibration of the vocal folds is the excitation source (Fant, 1981). This source-filter model of speech production is illustrated in figure 2.1.

It is convenient to describe the linear model of vowel production by

$$S(z) = E(z)G(z)V(z)R(z), \quad (2.1)$$

where $S(z)$ is the z-transform of the acoustic pressure at a given distance from the mouth. $E(z)$ is the z-transform of an impulse train that serves as the glottal excitation. $G(z)$ is the glottal shaping function which can be approximated by a second-order low-pass filter (Markel and Gray, 1976) of the form

$$G(z) = \frac{1}{(1 - e^{-cT}z^{-1})^2}, \quad (2.2)$$

where T is the sampling period and c is a positive real constant that controls the bandwidth of the low-pass filter. The glottal shaping function provides a skewed triangular airflow when the impulse train serves as the excitation (see Figure 2.2(a)). The glottal flow train is the combination of the glottal excitation $E(z)$ and the glottal shaping function $G(z)$. Its spectrum decreases at a rate of $12dB/octave$. (see Figure 2.3(a)).

$V(z)$ is the vocal tract transfer function which is modeled as an all-pole filter

$$V(z) = \frac{1}{\prod_{i=1}^K [1 - 2e^{-c_i T} \cos(b_i T) z^{-1} + e^{-2c_i T} z^{-2}]}, \quad (2.3)$$

where K is the number of formants or resonant frequencies. The i th formant frequency and bandwidth can be calculated from $F_i = b_i/2\pi$ and $B_i = c_i/2\pi$, respectively. The impulse response and frequency response of the vocal tract to produce the vowel /a:/ are shown in Figure 2.2(b) and Figure 2.3(b), respectively.

$R(z)$ is the lip radiation function, which can be approximated by a first-order high-pass filter, i.e.,

$$R(z) = 1 - z^{-1}. \quad (2.4)$$

When cT in equation 2.2 is much less than unity, one of the poles of the glottal function can be canceled by the zero of the radiation function. Thus, the entire transfer function $G(z)V(z)R(z)$, which is the combined spectral contributions of the

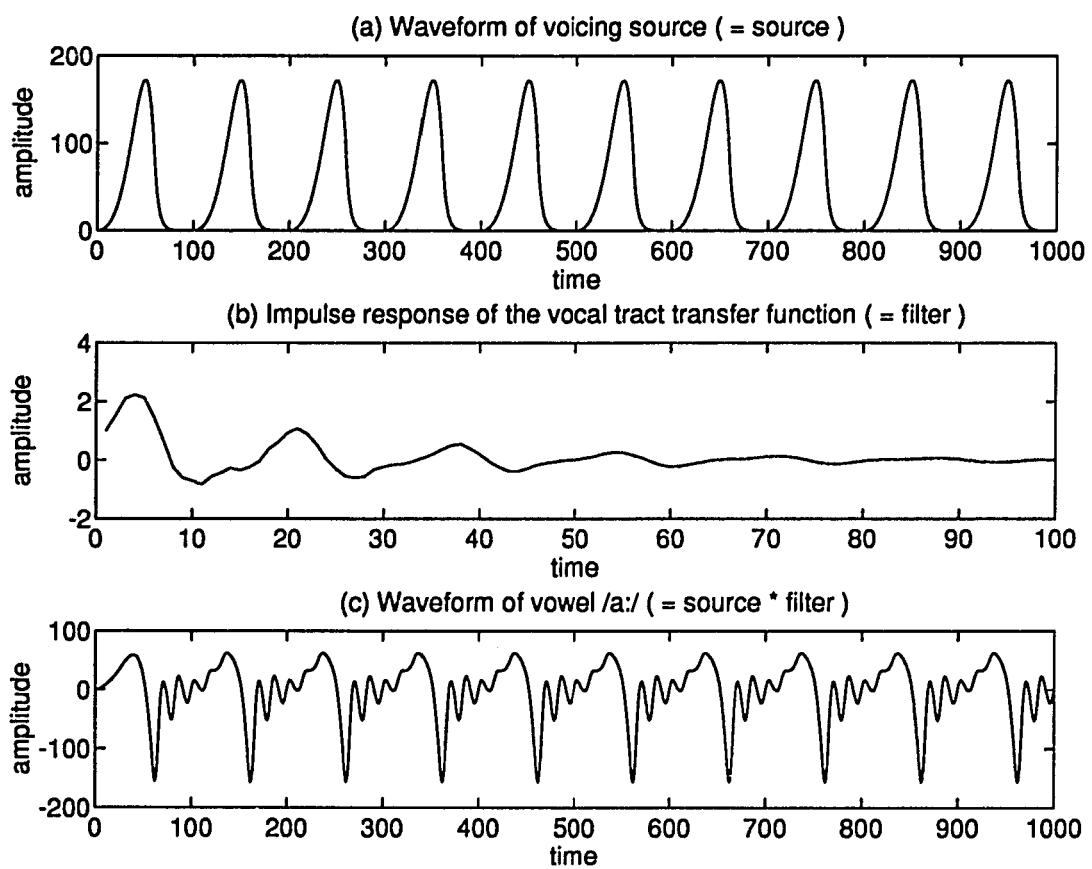


Figure 2.2: Temporal signal of source-filter model: (a) waveform of the voicing source, (b) impulse response of vocal tract, and (c) waveform of vowel

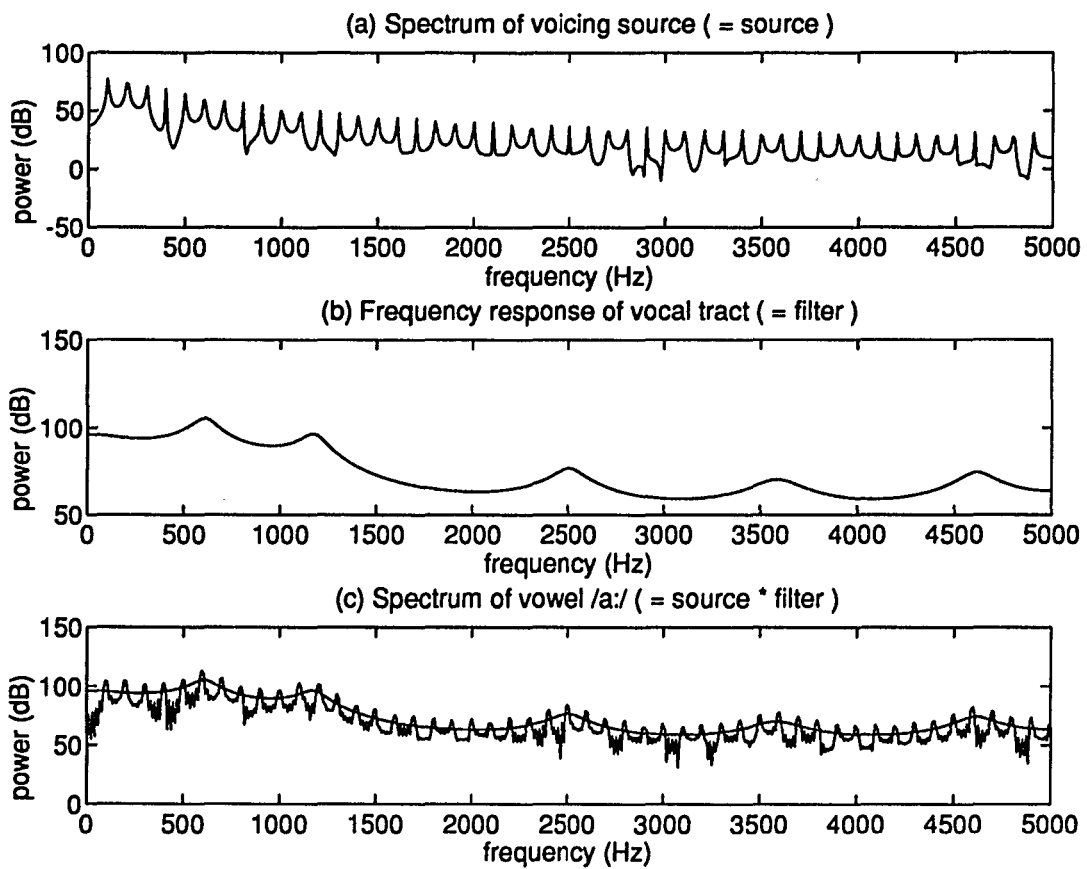


Figure 2.3: Spectrum of source-filter model: (a)spectrum of the voicing source, (b) frequency response of vocal tract, and (c) spectrum of vowel

glottal flow shape, the vocal tract, and the radiation of the lips, can be modeled as an all-pole filter $\frac{1}{A(z)}$:

$$S(z) = E(z)G(z)V(z)R(z) = E(z)\frac{1}{A(z)}, \quad (2.5)$$

where $A(z)$ is defined as

$$A(z) = 1 + \sum_{i=1}^M a_i z^{-i} \quad (2.6)$$

with $M \geq 2K + 1$.

Equation 2.5 is referred to as the source-filter synthesis model since the speech signal $S(z)$ (see Figure 2.2(c) and Figure 2.3(c)) will be the output of the all-pole filter $\frac{1}{A(z)}$ while the glottal excitation $E(z)$ is the input.

It has been indicated that the spectral properties of a speech signal are changed relatively slowly with time. Consequently, we can assume that the spectrum is fixed over time intervals on the order of 10 to 30msec, and, furthermore, a linear time invariant system can be used to describe the vocal tract transfer function at a speech instant. Normally, the properties of the system should be updated every 10msec. The speech analysis-synthesis procedure is based on this hypothesis.

Most of the recognized differences between alaryngeal speech and normal speech are related to the source of vowel production (Robbins et al., 1984; Trudeau and Qi, 1990). It should be possible to improve the voice quality by enhancing the alaryngeal speech voicing source. According to the source-filter theory of speech production, the spectral information, including the effects of vocal tract, glottal shaping, and

radiation, can be separated from the glottal excitation. By using a normal glottal excitation to replace the excitation of alaryngeal speech, it should be possible to produce a speech output with improved prosody. It has been shown that the linear predictive analysis-synthesis method has the ability to accomplish this task and improve the voice quality of alaryngeal speech (Qi, 1990; Qi et al., 1995).

2.2 Linear Predictive Analysis-Synthesis

The linear prediction techniques have been applied to the problem of modeling speech behavior for some time (Satio and Itakura, 1966; Atal and Schroeder, 1967), and have provided an effective mathematic tool to describe the linear source-filter model of speech production.

To multiply $A(z)$ to both sides of equation 2.5 results in the analysis model

$$E(z) = S(z)A(z). \quad (2.7)$$

(See Figure 2.4(a)). Its time domain expression can be

$$e(n) = s(n) + \sum_{i=1}^M a_i s(n-i) = \sum_{i=0}^M a_i s(n-i), \quad (2.8)$$

where $s(n)$ is the speech waveform signal, a_i is defined from equation 2.6 and $a_0 = 1$. For an M th order linear predictor, the predicted sample $\hat{s}(n)$ is determined by the previous M samples.

$$\hat{s}(n) = - \sum_{i=1}^M a_i s(n-i). \quad (2.9)$$

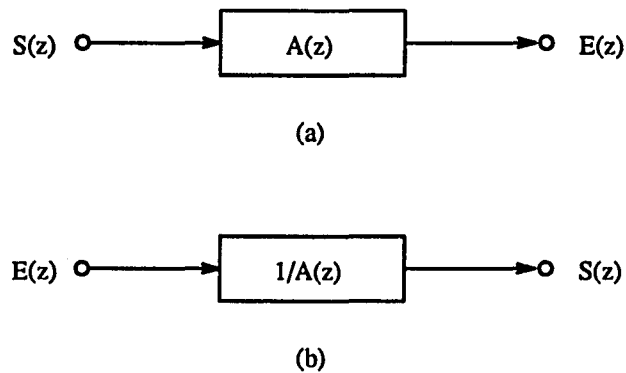


Figure 2.4: Linear prediction model: (a) analysis model, and (b) synthesis model

Then

$$e(n) = s(n) - \hat{s}(n). \quad (2.10)$$

Now, the $e(n)$ can be interpreted as the prediction error between the actual data sample $s(n)$ and the predicted sample $\hat{s}(n)$. The $a_i, i = 1, 2, \dots, M$, is defined as the predictor coefficients, which can be determined from the speech waveform by applying a least squares criterion to $e(n)$.

Let's define the mean square error, ϵ , as the sum of squares of the prediction errors over a range, n_0 to n_1 :

$$\epsilon = \sum_{n=n_0}^{n_1} e(n)^2 = \sum_{n=n_0}^{n_1} \sum_{i=0}^M \sum_{j=0}^M a_i s(n-i) s(n-j) a_j. \quad (2.11)$$

If defining the covariance function

$$\phi_{ij} = \sum_{n=n_0}^{n_1} s(n-i) s(n-j), \quad (2.12)$$

the equation 2.11 becomes

$$\epsilon = \sum_{i=0}^M \sum_{j=0}^M a_i \phi_{ij} a_j. \quad (2.13)$$

To minimize the mean square error by taking $\frac{\partial \epsilon}{\partial a_k} = 0$, for $1 \leq k \leq M$, we have

$$\sum_{i=0}^M a_i \phi_{ik} = 0, \quad (2.14)$$

or since $a_0 = 1$,

$$\sum_{i=1}^M a_i \phi_{ik} = -\phi_{0k}, \quad (2.15)$$

where $1 \leq k \leq M$. This is known as Yule-Walker equations. The predictor coefficients a_i , $1 \leq i \leq M$, can be obtained by solving this set of M linear simultaneous equations.

Two methods were developed to solve the Yule-Walker equations: the auto-correlation method and the covariance method. The auto-correlation method is defined by multiplying a finite length window $[0 \leq n \leq N - 1]$ to the speech signal $s(n)$. This limitation make the ϕ_{ij} as

$$\phi_{ik} = \sum_{n=0}^{N-1-(i-k)} s(n)s(n+i-k) = r(i-k), \quad (2.16)$$

where $r(k)$ is defined as the auto-correlation function. Because the auto-correlation function is symmetric, i.e. $r(k) = r(-k)$, the Yule-Walker equations can be expressed as

$$\sum_{i=1}^M a_i r(|i-k|) = -r(k), \quad 1 \leq j \leq M, \quad (2.17)$$

where

$$r(k) = \sum_{n=0}^{N-1-k} s(n)s(n+k), \quad k \geq 0,$$

with error

$$e(n) = \sum_{i=0}^M a_i s(n-i), \quad 0 \leq n \leq N+M-1.$$

The predictor coefficients a_i can be computed using Levinson-Durbin's recursion algorithm (Markel and Gray, 1976; Rabiner and Schafer, 1978):

1. The first step

$$k_1 = -a_1^{(1)} = r(1)/r(0), \quad (2.18)$$

$$\epsilon^{(1)} = (1 - k_1^2)r(0).$$

2. For iterations $i = 2, 3, \dots, M$

$$k_i = -a_i^{(i)} = [r(i) + \sum_{m=1}^{i-1} a_m^{(i-1)} r(i-m)]/\epsilon^{(i-1)},$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)}, \quad 1 \leq j \leq i-1.$$

$$\epsilon^{(i)} = (1 - k_i^2)\epsilon^{(i-1)}.$$

The parameters $k_i (i = 1, 2, \dots, M)$ are referred to as the reflection coefficients since they define the reflection properties of an multi-segment (with different cross segmentation area) acoustic tube model of the vocal tract.

The covariance method is defined by setting the $n_0 = M$ and $n_1 = N - 1$ in equation 2.11. So that the covariance function can be calculated by

$$\phi_{ij} = \sum_{n=M}^{N-1} s(n-i)s(n-j), \quad (2.19)$$

with error

$$e(n) = \sum_{i=0}^M a_i s(n-i), \quad M \leq n \leq N-1. \quad (2.20)$$

Since the set of ϕ_{ij} are known, equation 2.11 can be solved (Parsons, 1987).

In the linear predictive analysis procedure, a speech waveform is analyzed or coded by using an all-pole modeling method, which is also called the auto-regressive (AR) modeling method. The predictor coefficients a_i , the so-called linear predictive coding (LPC) coefficients, are used to describe the entire transfer function, including contributions of the glottal shaping, the vocal tract, and the radiation of lips, $\frac{1}{A(z)}$ (see equation 2.6). The prediction error signal $e(n)$ of this model, the so-called residual signal, is related to the voiced or unvoiced source.

According to the synthesis model (see equation 2.5), a speech waveform can be generated when an excitation signal is passed through the vocal tract filter (see Figure 2.4(b)). The synthesis process is usually implemented by two methods: the AR filter and the Lattice filter (Markel and Gray, 1976).

The AR filter is defined by an implementation of the direct form of the difference equation corresponding to the synthesis model. The equation 2.5 can be expressed

in time domain as

$$y(n) = e(n) + \sum_{i=1}^M a_i y(n-i), \quad (2.21)$$

where the output of the filter, $y(n)$, is the synthesized waveform; the input of the filter, $e(n)$, is the source excitation signal; and a_i is the predictor coefficients.

The Lattice filter is controlled by the reflection coefficients (see equation 2.18) based on the recurrence relations among prediction errors. The output of a Lattice filter is exactly the same as the output of an AR filter.

The predictor coefficients of the AR filter, or the reflection coefficients of the Lattice filter, are updated at the initiation of every pitch period sample in the excitation function $e(n)$. This procedure, referred to as pitch-synchronous synthesis, can minimize the effect of updating filter coefficients.

If $e(n)$ is the residual signal in equation 2.8, then $y(n)$ will be the original speech signal $s(n)$. Qi's method to improve the alaryngeal speech was to replace the residual signal by a synthetic source signal with higher and smoother fundamental frequency (Qi, 1990). In this work, the synthetic source signal is generated according to the LF model of glottal flow and its approximation (Fant et al., 1985; Qi and Bi, 1994).

2.3 Homomorphic Speech Processing

According to the linear model of speech production, the speech signal can be represented as a convolution of a source excitation and a filter impulse response. Therefore, the speech analysis can also be viewed as a problem of homomorphic

deconvolution (Rabiner and Schafer, 1978). The reasons we discuss this topic here are: 1. the homomorphic processing can also be used to implement decomposition and composition of the source and filter components for speech production; 2. the homomorphic processing would lead to an important parameter “cepstrum,” which is the coefficient we used to describe the vocal tract transfer function and to realize the spectral conversion in this work; 3. the homomorphic convolution was used to transfer the cepstrum to the impulse response of the vocal tract filter in our speech conversion system.

Assuming $f(n)$ is the impulse response of the vocal tract filter, and $e(n)$ is the excitation signal, the speech signal $s(n)$ is

$$s(n) = e(n) * f(n). \quad (2.22)$$

Implementing a z-transform of the equation 2.22, the convolution between the $e(n)$ and the $f(n)$ then becomes a product between the $E(z)$ and the $F(z)$:

$$S(z) = E(z) \cdot F(z), \quad (2.23)$$

where $S(z) = Z[s(n)]$, $E(z) = Z[e(n)]$, and $F(z) = Z[f(n)]$. Since the logarithms of a product is equal to the sum of the logarithms of the individual terms, we have

$$\hat{S}(z) = \log[S(z)] = \log[E(z) \cdot F(z)] = \log[E(z)] + \log[F(z)]. \quad (2.24)$$

The inverse z-transform of $\hat{S}(z)$

$$\hat{s}(n) = Z^{-1}[\log[E(z)] + \log[F(z)]] = Z^{-1}[\log[E(z)]] + Z^{-1}[\log[F(z)]]. \quad (2.25)$$

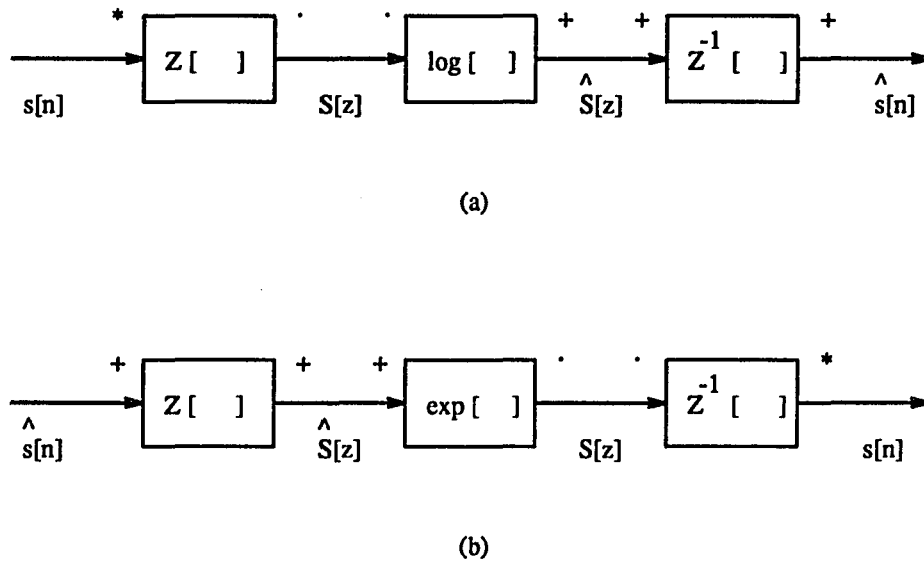


Figure 2.5: Homomorphic processing: (a) homomorphic deconvolution, and (b) its inverse process

From equation 2.25, it is clear that the combination in the manner of convolution between the $e(n)$ and the $f(n)$ has been represented as an additive combination in $\hat{s}(n)$. This process is called the homomorphic deconvolution and is shown in Figure 2.5(a). The inverse of this process is shown in Figure 2.5(b).

The $F(z)$ in equation 2.23 is actually the $\frac{1}{A(z)}$ in equation 2.5 for vowel production.

If including a gain factor G in here, we have

$$F(z) = G \frac{1}{A(z)} = G \frac{1}{1 + \sum_{i=1}^M a_i z^{-i}} = G \frac{1}{\prod_{k=1}^M (1 - p_k z^{-1})}, \quad (2.26)$$

where p_k corresponds to poles inside the unit circle in z -plan. The logarithms of the $F(z)$ is

$$\log[F(z)] = \log(G) - \sum_{k=1}^M \log(1 - p_k z^{-1}). \quad (2.27)$$

Using the power series expansions of $\log(1 - \alpha z^{-1}) = -\sum_{n=1}^{\infty} \frac{\alpha^n}{n} z^{-n}$, for $|\alpha| < |z|$, we have

$$Z^{-1}[\log[F(z)]] = \begin{cases} \log(G) & n = 0 \\ \sum_{k=1}^M \frac{p_k^n}{n} & n > 0 \\ 0 & n < 0 \end{cases} \quad (2.28)$$

From equation 2.28, it is clear that the portion in the $\hat{s}(n)$ related to the vocal tract, glottal pulse shape, and radiation information decays at least as fast as $1/n$. So, it is reasonable to assume that the low-time part of the $\hat{s}(n)$ corresponds primarily to the vocal tract, glottal pulse, and radiation information, while the high-time part is due primarily to the excitation. Using a window to select the low-time part of the $\hat{s}(n)$, the homomorphic decomposition of the source and filter components for speech production is realized.

For computational purpose we use the Fourier transform to replace the z-transform in equation 2.25, i.e., let $z = e^{j\omega}$, then

$$\hat{s}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log[S(e^{j\omega})] e^{j\omega n} d\omega = \frac{1}{2\pi} \int_{-\pi}^{\pi} [\log |S(e^{j\omega})| + j \arg S(e^{j\omega})] e^{j\omega n} d\omega. \quad (2.29)$$

It is called the complex cepstrum. The cepstrum is defined as the inverse Fourier transform of the logarithm of the power spectrum of a signal:

$$c(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |S(e^{j\omega})| e^{j\omega n} d\omega. \quad (2.30)$$

By comparing equation 2.32 and equation 2.30, we can see that $c(n)$ is the inverse transform of the real part of $\hat{S}(e^{j\omega})$. Therefore, $c(n)$ is equal to the conjugate-symmetric part of $\hat{s}(n)$ (Oppenheim and Schaffer, 1989), that is

$$c(n) = \frac{\hat{s}(n) + \hat{s}(-n)}{2}. \quad (2.31)$$

The $\hat{s}(n)$ can be recovered from $c(n)$ by frequency-invariant linear filtering if $\hat{s}(n)$ is causal, i.e., $s(n)$ is minimum phase,

$$\hat{s}(n) = c(n)l_{min}(n), \quad (2.32)$$

where $l_{min}(n) = 2[u(n) - u(n - N)] - \delta(n)$, it is called the cepstrum window with window length N . The cepstral coefficients are truncated to limited length. The truncated cepstral coefficients draw a smoothed curve to describe the log magnitude of the spectrum.

Furthermore, if we implement an inverse process of homomorphic deconvolution to the $\hat{s}(n)$, we will obtain a minimum-phase sequence $s_{min}(n)$ such that $|S_{min}(e^{j\omega})| = |S(e^{j\omega})|$. Obviously, the $s_{min}(n)$ is equal to the impulse response of the filter in the linear system of speech production. The basic strategy of the homomorphic filtering to extract the cepstrum and the impulse response of the filter in the linear system of speech production is shown in the block diagram of Figure 2.6.

It has been revealed that the L_2 norm distance measure of the cepstrum is very close to that of the log spectrum according to the Parseval's relation (Gray and

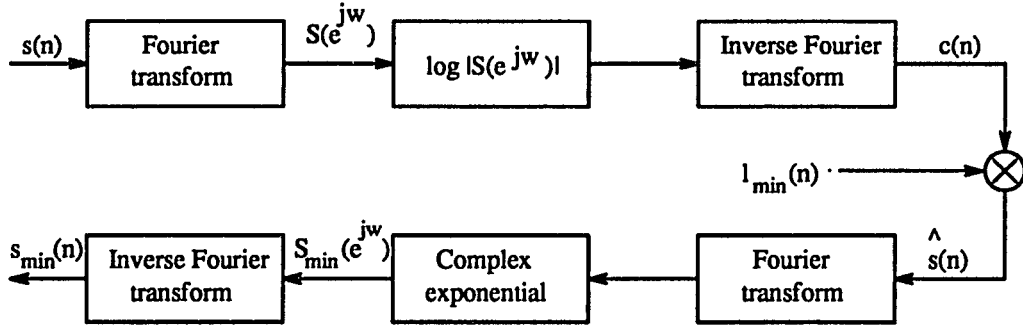


Figure 2.6: The use of homomorphic filtering to extract the cepstrum and the impulse response of the system

Markel, 1976). Therefore, we use the cepstral coefficients as the parameters to measure the distance among spectra in the spectral conversion process.

Both the LPC coefficients and the cepstral coefficients are used to describe the spectral information. The relationship between the LPC coefficients and the cepstral coefficients can be found from the definition of cepstrum, i.e., the z-transform of cepstrum equals to the logarithm spectrum,

$$\log\left[\frac{1}{A(z)}\right] = \sum_{i=1}^{\infty} c_i z^{-i}, \quad (2.33)$$

where $A(z)$ has been defined by equation 2.6 and c_i is the same of $c(i)$ in previous equations. By differentiating both sides of equation 2.33 with respect z^{-1} , we have

$$\sum_{i=1}^M i a_i z^{-i} = -\left[\sum_{i=1}^{\infty} i c_i z^{-i}\right] \left[\sum_{i=0}^M a_i z^{-i}\right]. \quad (2.34)$$

Considering $a_0 = 1$, we have the recursion relations

$$c_1 = -a_1, \quad (2.35)$$

$$jc_j = -ja_j - \sum_{i=1}^{j-1} ic_i a_{j-i}, \quad j = 2, 3, \dots, M,$$

$$jc_j = \sum_{i=1}^M (j-i)c_{j-i}a_i, \quad j = M+1, M+2, \dots$$

To include the gain constant of the LPC spectral model in the set of cepstral coefficients, c_0 is defined as the normalized prediction error.

The complex cepstral coefficient (c_j) can be considered as the j th-order coefficient of a Taylor series expansion of the $\log[\frac{1}{A(z)}]$ (Gray and Markel, 1976). Superficially, infinite cepstral coefficients are needed to represent the logarithm spectrum. It has been shown, however, that the first N cepstral coefficients will adequately represent the logarithm spectrum. Therefore, the cepstrum computed from the LPC coefficients is a N th-order approximation of the logarithm all-pole model $\log[\frac{1}{A(z)}]$.

So far, we have had three methods to extract the vocal tract transfer function from speech waveform: the all-pole model by LPC coefficients, the smoothed log magnitude function by cepstral coefficients, and the approximation of the all-pole model by cepstral coefficients calculated from the LPC coefficients, which is referred to as the LPC cepstrum. Three curves and the Fourier transform of the vowel /a:/ are shown in Figure 2.7. All of the three curves can be used to describe the envelope of the power spectrum.

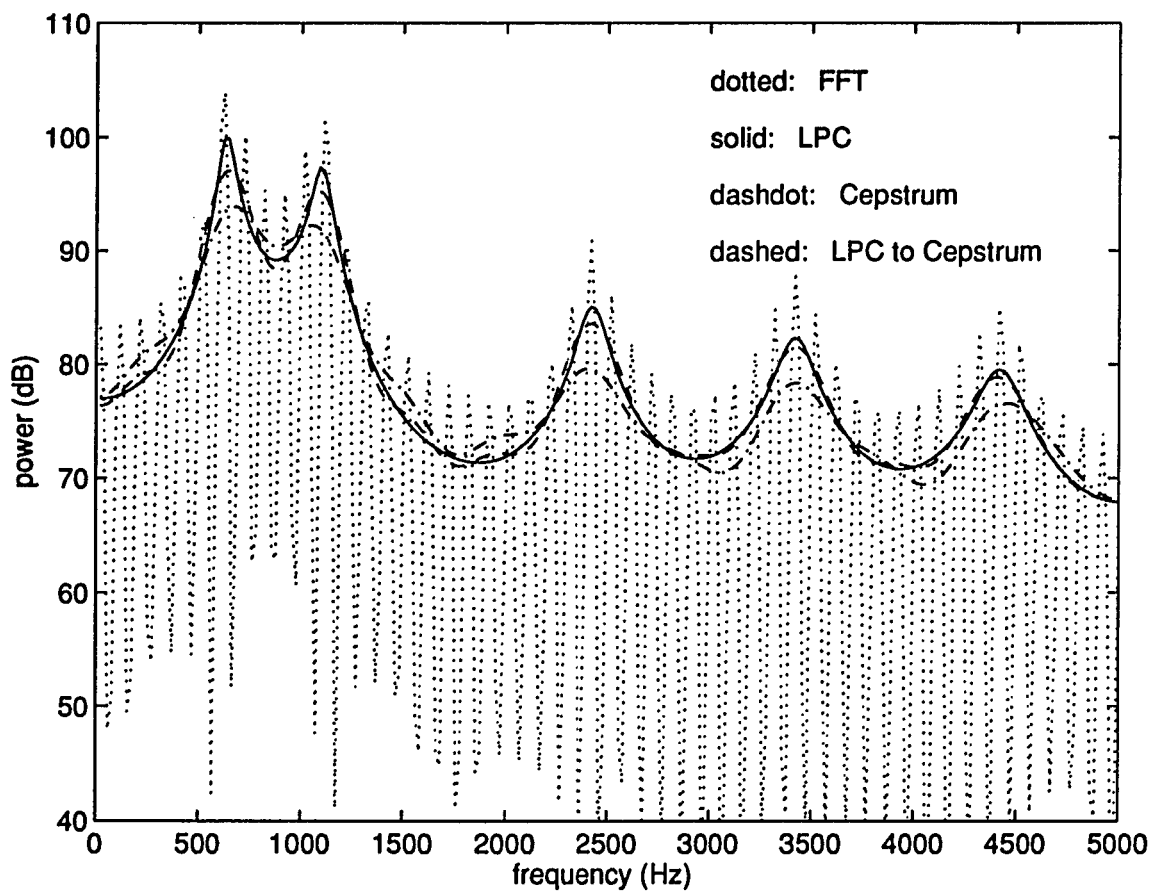


Figure 2.7: Spectra of LPC coefficients, cepstral coefficients, and LPC cepstral coefficients

2.4 Vowel Detection and Pitch Period Estimation

The excitation source for voiced and unvoiced speech is totally different. Therefore, a vowel detector is important to any analysis-synthesis system. The cepstrum calculated from the homomorphic deconvolution provides a way to distinguish between voiced and unvoiced speech. Because of the periodicity of voiced speech, there is a peak in the cepstrum at the fundamental period of the input signal. An input speech segment is likely to be voiced if its cepstrum peak exceeds a pre-set threshold; the position of the peak is the pitch period. An input speech segment is unvoiced if the peak is not above the threshold (Noll, 1967).

If c_i denotes the cepstral coefficient, the cepstrum peak value, which is the voicing determination parameter, is given by

$$P = \max_{n_1 \leq i \leq n_2} [c_i], \quad (2.36)$$

where n_1 to n_2 is the peak searching range. The threshold of the cepstrum peak for voicing determination of normal speech can be 0.1. The zero-crossing rate and energy are also used in making this decision.

2.5 Glottal Flow Model and Its Approximation

An suitable model of the human voice source is essential for synthesizing natural sounding speech (Klatt and Klatt, 1990; Childers and Lee, 1991). The 4-parameter model developed by Liljencrants, Fant and Lin (referred to as the LF model) has been

widely accepted as an effective source model (Fant et al., 1985). In the LF model, the flow derivative of the voice source is specified by 4 temporal parameters, t_c , t_p , t_e , and t_a (see Figure 2.8). Three of the parameters have direct physical correspondence with human voicing events: t_c could be the fundamental period, t_p is the instant of maximum flow, and t_e is the instant of maximum glottal closing speed, which is also the minimum of the flow's derivative. Parameter t_a is the second (right) derivative of the volume flow at the minimum of the first derivative. It does not have apparent physical correspondence with human voicing events. These parameters are referred to as the analysis set of parameters and they may be estimated from inverse filtering of the speech signal.

The LF model is specified mathematically using several additional parameters, E_0 , E_e , ω_g , α , and ϵ . These parameters are used to synthesize the flow derivative in the LF model.

$$E(t) = \begin{cases} E_1(t) = E_0 e^{\alpha t} \sin \omega_g t & (t \leq t_e), \\ E_2(t) = \frac{-E_e}{\epsilon t_a} [e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}] & (t_e < t \leq t_c), \end{cases} \quad (2.37)$$

where $E(t)$ is the flow derivative.

All parameters in equation 2.37 need to be supplied explicitly in any numerical implementation of the LF model. Some parameters are relatively easy to obtain. For example, E_0 is an arbitrary gain constant and ω_g simply equals $\frac{\pi}{t_p}$. Others are not. For example, we have to solve the following exponential equation to get the explicit

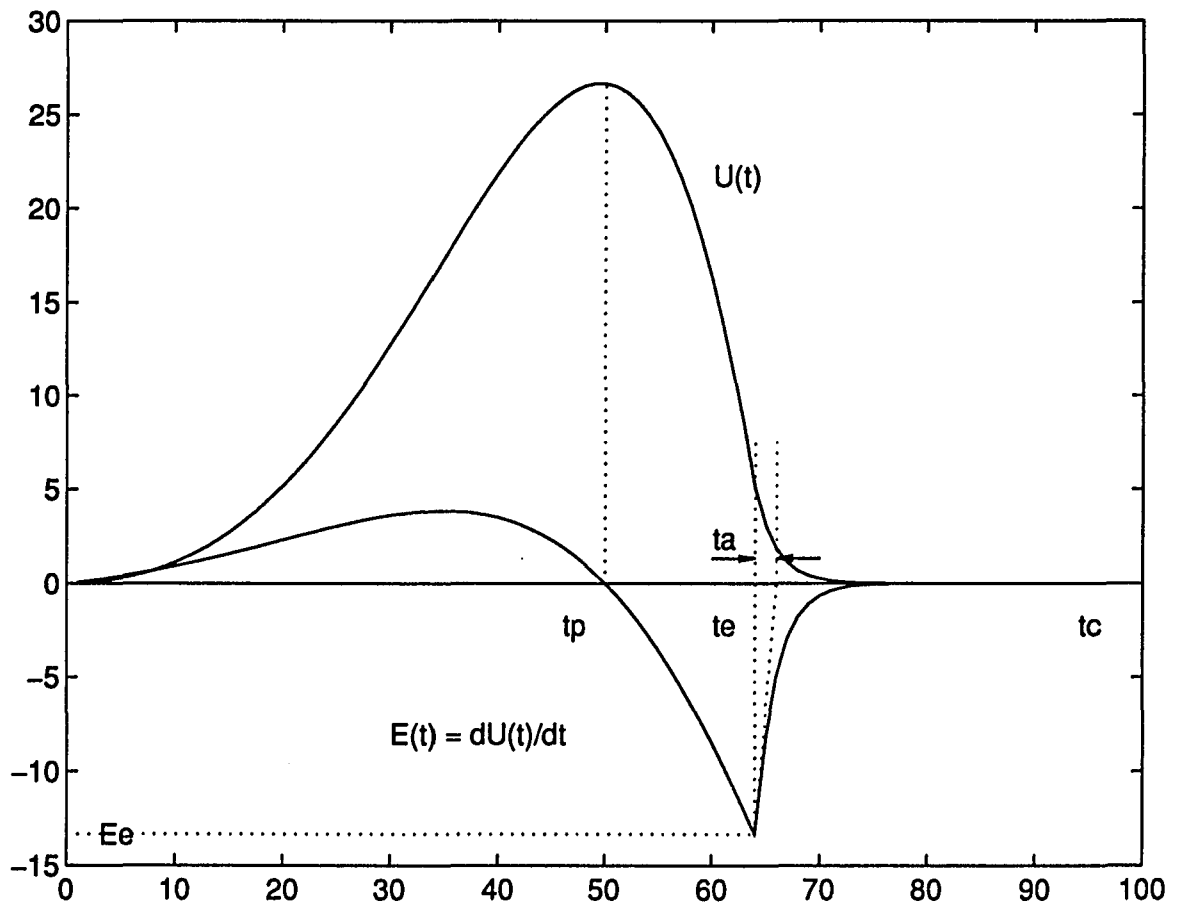


Figure 2.8: LF model of glottal flow

value of ϵ

$$1 - e^{-\epsilon(t_c - t_e)} = \epsilon t_a. \quad (2.38)$$

Equation 2.38 results from setting $E_2(t_e) = -E_e$, where $E_e > 0$ is the amplitude of flow derivative at t_e . Once the ϵ is obtained, the α can be calculated by solving the following integral equation

$$\int_0^{t_c} E(t) dt = 0. \quad (2.39)$$

Equation 2.39 ensures that the source flow function returns to baseline at the end of each period. Finally, E_e can be computed as

$$E_e = -E_0 e^{\alpha t_e} \sin(\omega_g t_e). \quad (2.40)$$

The model is computationally complex since it requires the solution of roots of a nonlinear equation and an integral equation for each given set of model parameters.

The complicated relationships among parameters of the LF model are largely related to the constraint imposed by equation 2.39. This constraint ensures that both the flow and the flow derivative return to the baseline at the end of each period. This constraint also forces all parameter adjustments to be made concurrently on both segments of the LF model and imposes complicated relationships among model parameters.

One simplified approximation was discussed recently (Qi and Bi, 1994). Instead of the t_a in the second segment of the LF model, E_e is used as an explicit model

parameter (Ananthapadmanabha and Fant, 1982; Ananthapadmanabha, 1984). As a result, the 4 parameters of the voice source model now include t_c , t_p , t_e , and E_e . In this approximation, the first segment of the LF model remains unchanged. The second segment of the LF model remains an exponential function, but takes a slightly different form

$$E_2(t) = -E_e e^{-\epsilon(t-t_e)}, \quad (2.41)$$

where E_e is a known constant. The antiderivative of this equation is

$$U_2(t) = \int_{t_e}^t E_2(t) dt = \frac{E_e}{\epsilon} e^{-\epsilon(t-t_e)} - \frac{E_e}{\epsilon} e^{-\epsilon(t_e-t_e)}. \quad (2.42)$$

Since E_e is given, using equations 2.40, α is simply

$$\alpha = \frac{1}{t_e} \ln\left(\frac{-E_e}{E_0 \sin(\omega_g t_e)}\right) \quad (2.43)$$

ϵ is obtained by the flow continuity constraint,

$$U_1(t_e) = U_2(t_e), \quad (2.44)$$

where

$$U_1(t) = \int_0^t E_1(t) dt = \frac{e^{\alpha t} (\alpha \sin(\omega_g t) - \omega_g \cos(\omega_g t)) + \omega_g}{\alpha^2 + \omega_g^2}, \quad (2.45)$$

and $U_2(t)$ is shown in equation 2.42. Since the α can be computed directly from equation 2.43, the root of only one non-linear equation (equation 2.44) needs to be solved to compute the source function. This is referred to as the approximation I of the LF model. If it is further assumed that the closed phase is relatively long and the return to flow baseline is relatively fast following glottal closing (i.e., $\epsilon(t_c - t_e) \gg 1$),

the term $e^{-\epsilon(t_c-t_e)}$ would be near zero. Under these assumptions, ϵ can be directly obtained as

$$\epsilon = \frac{E_e}{U_1(t_e)}, \quad (2.46)$$

where $U_1(t)$ is not a function of ϵ (see equation 2.45). No root solving is necessary.

This is referred to as the approximation II of the LF model.

The flow and its derivative under various parameter values for approximation I and II are shown in Figure 2.9, respectively. Here t_c , t_e and t_p were held constant and E_e was varied from -5 to -15. The flow and flow derivative of approximation I and II were nearly identical to those of the LF model. It has been shown that the spectral range of two approximation models were also closed to that of the LF model (Qi and Bi, 1994). The results demonstrated that this simplified approximations are capable of producing a wide range of source functions characterized by temporal and spectral content that is comparable to that produced by the LF model.

The approximation II of LF model was adopted as the voicing source signal for speech synthesis in this work.

2.6 Dynamic Time Warping in Pattern Alignment

Dynamic time warping (DTW) algorithm is a method used to achieve alignment of two patterns that will be used in this work to align the same phonemic events of two utterances. Its basic principles are summarized here (Nemhauser, 1966; Itakura, 1965; White and Neely, 1976; Rabiner et al., 1978; Parsons, 1987).

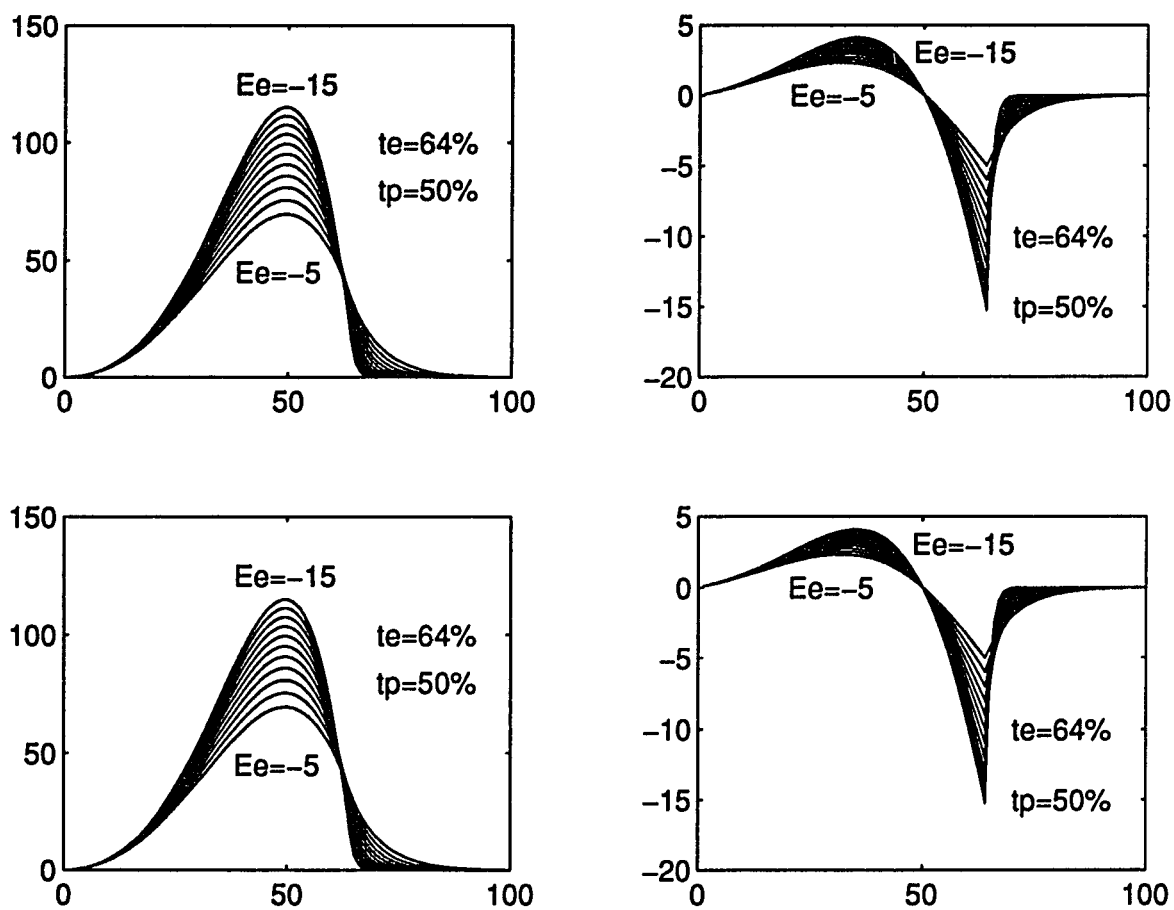


Figure 2.9: Approximation of LF model: (a) flow of approximation I, (b) flow derivative of approximation I, (c) flow of approximation II, and (d) flow derivative of approximation II

The inputs to the DTW process are two patterns with series of vectors, i.e., the LPC cepstra of the same word pronounced by two speakers, and the outputs are a warping function and the degree of remaining mismatch. The warping function provides a path for the optimal match between the inputs. The remaining mismatch, referred to as the “total cost” in dynamic programming terminology, reflects the inherent discrepancies between the two inputs.

Suppose that two patterns to be matched are

$$A = a_1, a_2, \dots, a_i, \dots, a_M, \quad (2.47)$$

$$B = b_1, b_2, \dots, b_j, \dots, b_N, \quad (2.48)$$

where a_i and b_j are the i th element of pattern A and the j th element of B , respectively. Computation of the warping function can be viewed as the process of finding a minimum-cost path through the lattice of points depicted in Figure 2.10.

Let C be the warping function

$$C = c_1, c_2, \dots, c_k, \dots, c_L, \quad (2.49)$$

where $c_k = [i_k, j_k]$ is a pair of pointers to the samples being matched. For each c_k , the cost can be calculated by the Euclidean distance as

$$d(c_k) = (a_{i_k} - b_{j_k})^2. \quad (2.50)$$

The final warping function is determined by minimizing the total-cost function

$$D(C) = \sum_{k=1}^L d(c_k) \quad (2.51)$$

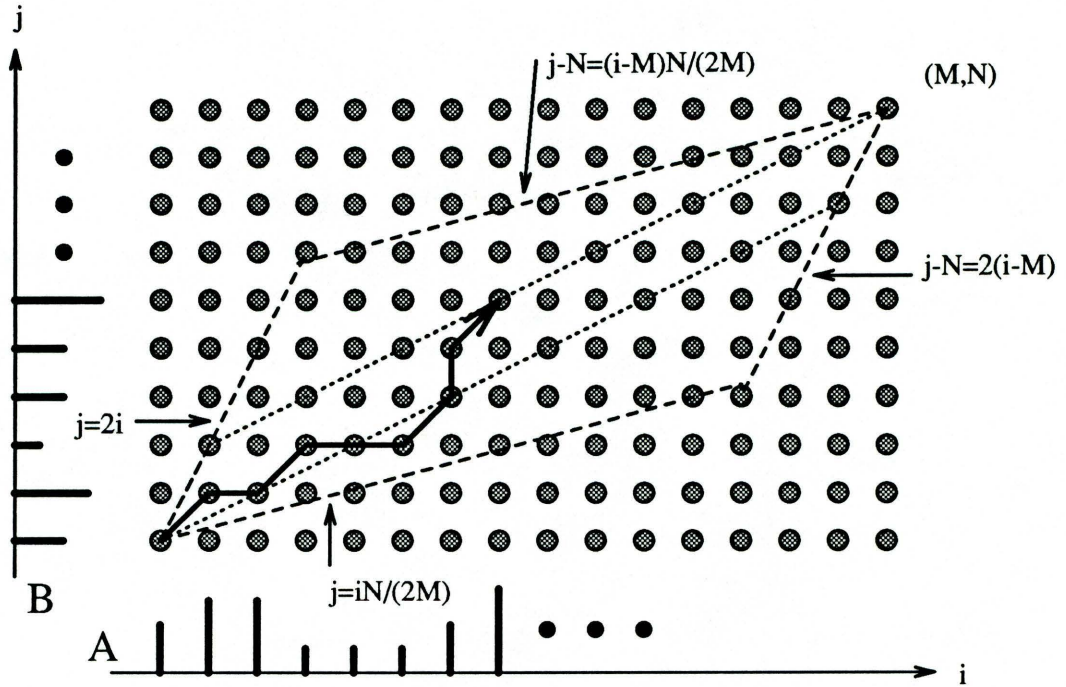


Figure 2.10: A schematic illustration of dynamic programming

subject to monotonic constraints

$$i_k \geq i_{k-1}, j_k \geq j_{k-1}, \quad (2.52)$$

continuous constraints

$$i_k - i_{k-1} \leq 1, j_k - j_{k-1} \leq 1, \quad (2.53)$$

and some kind of global limit of the maximum amount of warp

$$|i_k - j_k| < W, \quad (2.54)$$

where W is defined as the window width that describes a region where the warp is allowed. Usually, the global limit can be imposed on the slope of the warping function (the dotted line shown in Figure 2.10) where the slopes of the parallelogram's sides

are 1 to 2 and 2 to 1, respectively. This parallelogram method works well if M and N are roughly equal because the parallelogram will collapse into a line and no warping can be done if $M \geq 2N$ or $N \geq 2M$.

2.7 Voicing Source Replacement and Synthesis

The synthetic excitation source for voiced speech should be an impulse train if the spectrum contains $-6dB/octave$ trend, which is a combination of a $-12dB/octave$ trend due to the glottal shaping function and $+6dB/octave$ trend due to radiation from the mouth. In speech analysis processing, a pre-emphasis of the input signal is usually employed to give a $+6dB/octave$ lift in the appropriate range so that the measured spectrum has a similar dynamic range across the entire frequency band. The pre-emphasis can be implemented as a first-order high-pass filter with transfer function

$$H(z) = 1 - az^{-1}, \quad (2.55)$$

where a is a constant usually chosen between 0.9 and 1. In time domain, it is a difference equation of

$$y(n) = x(n) - ax(n - 1), \quad (2.56)$$

where the $x(n)$ denotes the current input sample, the $x(n - 1)$ denotes the previous input sample, and the $y(n)$ is the current output sample of the pre-emphasis filter.

One pole in the glottal shaping function (equation 2.2) will be canceled by the zero of

the pre-emphasis function (equation 2.55), while the other pole in the glottal shaping function has been canceled by the zero of the radiation function (equation 2.4).

If the LPC or cepstral coefficients used to describe the vocal tract transfer function are computed from the pre-emphasized data, the derivative of glottal flow model can be used to generate an overall -6dB/octave spectral trend in the synthetic speech. It has been shown that the LF model of the human voice source is able to achieve natural sounding synthetic speech (Childers and Lee, 1991). We would use the approximation of LF model as the excitation source in this work.

From discussion in previous sections, it is clear that the excitation sources can be separated from the vocal tract transfer functions by the LPC or the homomorphic processing. Therefore, we can use synthetic excitation sources to replace the alaryngeal excitation sources while retaining a speaker's vocal tract transfer functions to build a synthetic speech using the LPC analysis-synthesis, or the homomorphic deconvolution and its inverse processing, or a combination of these two methods.

Qi (1990, 1995) demonstrated that the LPC method can reliably extract vocal tract transfer functions of alaryngeal vowels despite the presence of large perturbations in fundamental frequency. Speech synthesized with a reconstructed transfer function and a synthetic excitation were shown to be intelligible and have improved source-related properties over those present in the original alaryngeal vowels. The basic strategy of the alaryngeal speech enhancement by LPC analysis-synthesis is shown in the block diagram of Figure 2.11. In this procedure, the pitch-synchronous

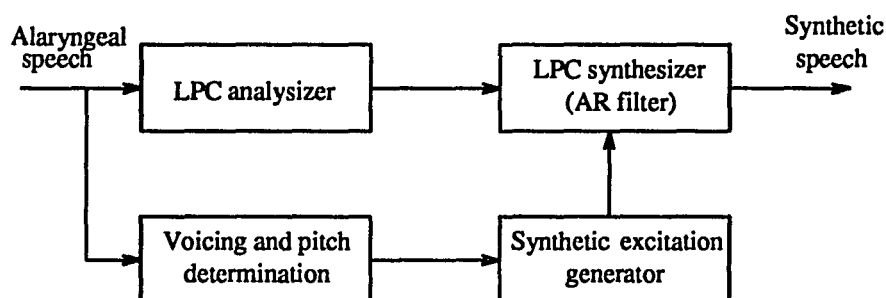


Figure 2.11: Block diagram of the LPC analysis-synthesis system to improve alaryngeal speech

LPC analysis and synthesis was used. LPC coefficients were computed for each voiced segment (or frame) using the auto-correlation method (see equation 2.18). Hamming window and pre-emphasis were used in the LPC analysis. The window length was established at 40 msec. The frame step-size was set to equal the current pitch period. The voicing determination and pitch period estimation were achieved by the peak evaluation of

the auto-correlation function. The synthetic voice source signal was generated from the LF model or its approximation. The AR filter was used to convolve the synthetic excitation and the LPC coefficients (see equation 2.21).

The homomorphic processing to enhance alaryngeal speech is the second procedure shown in Figure 2.12. In this case, the smoothed logarithm spectrum was estimated from the homomorphic deconvolution (see equation 2.30 and 2.32). Hamming window and pre-emphasis were implemented. The window length and step-size were the same as in the LPC procedure. The voicing detection and the pitch period estimation

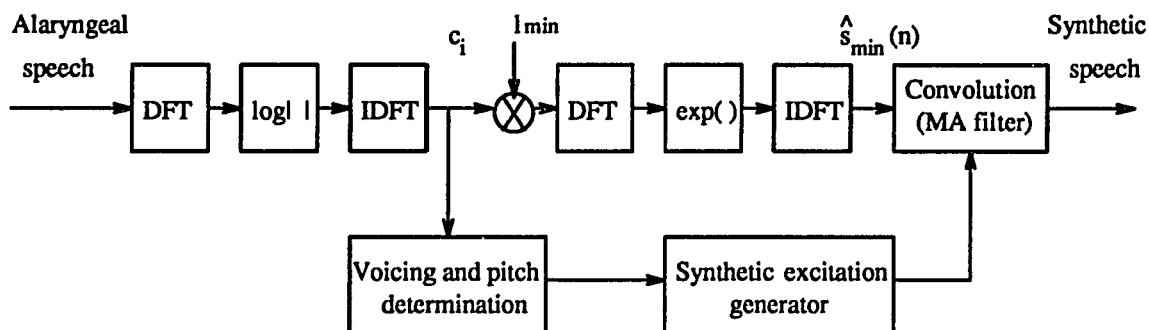


Figure 2.12: Block diagram of the homomorphic processing to improve alaryngeal speech

were achieved by implementing the cepstrum peak evaluation (see equation 2.36). During synthesis, the impulse response of the vocal tract filter was convolved with the synthetic voice source signal by driving a moving average (MA) filter.

If the LPC method is used to generate the cepstral coefficients, we can construct the third procedure to improve alaryngeal speech (see Figure 2.13). From the equation 2.35, it is clear that this process is an approximation of the first procedure. Furthermore, by using an all-pole model to estimate the frequency response of the minimum phase signal $\hat{s}_{min}(n)$, we would be able to build synthetic speech by the AR filter instead of the MA filter.

Speech synthesized by using the first and the third procedures shows little difference because both of their vocal tract transfer functions are estimated based on the all-pole model. Speech synthesized by using the second procedure is not as smooth as those synthesized by using other methods.

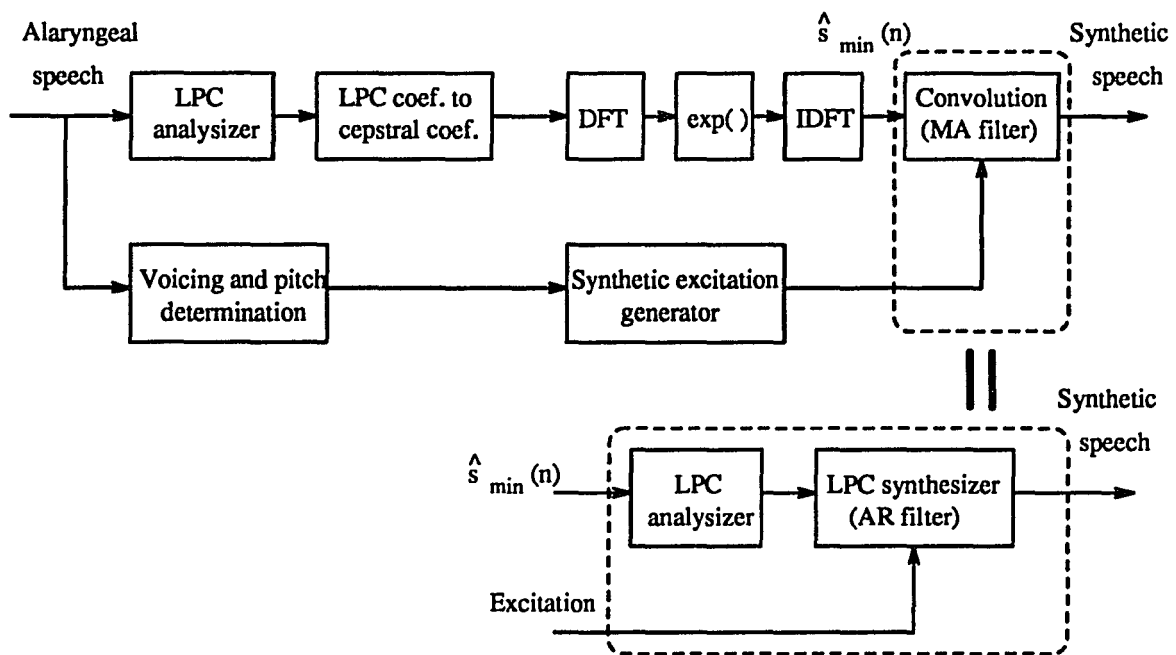


Figure 2.13: Block diagram of the combination of LPC and homomorphic processing to improve alaryngeal speech

The reason we introduced the third procedure is that the spectral conversion, which will be the main topic of the next chapter, can be realized in the form of cepstrum because the cepstral coefficients are a Euclidean space feature while the LPC coefficients are not. The cepstrum will work as an intermediate parameter to build a relationship between the spectral feature space of an alaryngeal speaker and that of a normal speaker.

CHAPTER 3

SPEECH CONVERSION

Current speech conversion algorithms were mainly designed for normal speech conversion (Abe et al., 1988; Shikano et al., 1991; Savic and Nam, 1991; Valbret et al., 1992). These algorithms usually were evaluated on the basis of whether the converted speech sounded closer to the target than to the original speech. The quality of the converted speech was not a major concern. To use speech conversion techniques for speech enhancement, the quality of converted speech was the primary concern. In this chapter, the vector quantization (VQ)-based and the linear multivariate regression (LMR)-based conversion methods were evaluated using simulated data. The purpose of the evaluation was to reveal the pros and cons of each method and to identify aspects of these methods where improvement could be made.

Both speech conversion algorithms consists of two phases: the learning phase and the conversion phase. In the learning phase, a conversion mapping function was generated based on given input and target vectors. In the conversion phase, an input vector is transformed into an output vector based on the mapping functions generated in the learning phase. Although the conversion processes are similar between the VQ- and LMR-based algorithms, the generation of mapping function is fundamentally different.

In the VQ-based approach, the learning is to generate an input codebook and a mapping codebook. The input codebook typically is generated by using vector quantization algorithms. These algorithms produce a set of representative vectors (codewords) by minimizing the total differences between the codewords and all learning input vectors. The mapping codebook specifies the output vector of an input codeword. This output vector usually is obtained by averaging the target vectors projected from a cluster of input vectors that belong to the same input codeword.

Two factors are crucial to the VQ-based learning: the size of input codebook and the number of learning pairs (input and target vectors). The larger the input codebook size, the more accurate the codebook will be. An unlimited codebook size, however, is neither practical nor desirable considering the existence of outliers or idiosyncratic variations in the input space. Likewise, the larger the number of learning pairs, the more accurate the resulting codebooks will be. The number of learning pairs often is limited.

The LMR-based approach makes continuous mapping between the input and output feature spaces, eliminating quantization distortions that are intrinsic to the discrete, VQ-based method. In the LMR-based approach, the learning is to generate a mapping matrix that linearly transform an input vector to an output vector. When the transformation between input and output feature spaces is approximately linear, the mapping matrix can be easily obtained using least-square approximations. If the transformation between input and output feature spaces is not linear, more than one

mapping matrix is necessary. The number of mapping matrices is a crucial factor in LMR-based conversion. A small number of mapping matrices may compromise the inherent non-linearity between the input and output feature spaces. A large number of mapping matrices, however, may degenerate the LMR-based approach to the VQ-based approach.

The accuracy of LMR-based conversion is also limited by the number of learning pairs. There has to be a sufficient number of input and target vector pairs to derive a non-singular mapping matrix. When the number of learning pairs is limited, some mapping matrices might be singular, and the mapping result might be unpredictable.

To appreciate which method could make optimal use of available information for achieving accurate conversion, and how to improve the available conversion methods, the VQ- and LMR-based algorithms were evaluated. Evaluation results, together with the details of VQ- and LMR-based conversion algorithms and their variations, are presented as follows.

3.1 VQ-Based Spectral Conversion

The VQ-based conversion system consists of two major components: an input codebook and a mapping codebook. The input codebook typically is obtained using vector quantization algorithms. The mapping codebook needs to be generated through a supervised learning process.

3.1.1 Vector Quantization and Fuzzy Vector Quantization

The vector quantization is a process to create an optimal codebook such that the total distortion between the codewords and the group of training vectors is minimized. Assuming \vec{c}_n is the n th codeword of the codebook with size N and \vec{v}_m is m th vector of the training set with size M , the goal of VQ is to select the set of codeword \vec{c}_n such that the total distortion between all of the training vectors, $\vec{v}_m, m = 1, \dots, M$ and codewords, $\vec{c}_n, n = 1, \dots, N$, is minimized. This total distortion can be defined as (Rabiner et al., 1983)

$$\|D_N\| = \min_{\vec{c}_n} \left\{ \frac{1}{M} \sum_{m=1}^M \min_{1 \leq n \leq N} [d(\vec{c}_n, \vec{v}_m)] \right\} \quad (3.1)$$

where $d(\vec{c}_n, \vec{v}_m)$ is the distance between two vectors. An effective algorithm for vector quantization has long been developed (Linde et al., 1980), and is often referred to as the LBG algorithm.

To visualize and evaluate the VQ process, 1000 random samples were generated in a two-dimensional squared space defined by $[(-5, -5), (5, 5)]$. A codebook with 128 codewords was generated using the LBG algorithm (see Figure 3.1(a)). It can be seen that the codebook is a reasonable representation of these samples.

Once the codebook is established, any given input vector is represented or encoded by its nearest codeword. Obviously, the encoding will introduce distortions. A demonstration of VQ distortion is shown in Figure 3.1(b)) when using the codebook

to encode 16 samples in a circle. It can be seen that the encoded circle is no longer a perfect circle.

The distortion of VQ processing cannot be avoided, but can be minimized. One way to reduce the VQ distortion is to increase the size of codebook. The downside of this method is that it requires more training vectors and computations. An alternative is to use the fuzzy vector quantization (FVQ) technique (Tseng et al., 1987; Nakamura and Shikano, 1989; Shikano et al., 1991). In FVQ, an input vector was coded as a weighted linear combination of k -nearest codewords. The weight is determined by a fuzzy membership function. For example, an input vector \vec{v} is encoded as

$$\vec{v} \rightarrow \sum_{i=1}^k f_i \vec{C}_i, \quad (3.2)$$

where k is the number of participating codewords. \vec{C}_i is the i th nearest codeword for the given input vector, and f_i is the membership of \vec{v} in cluster i . The membership function is defined as,

$$f_i = \frac{1}{d(\vec{C}_i, \vec{v})} / \sum_{l=1}^k \frac{1}{d(\vec{C}_l, \vec{v})}, \quad (3.3)$$

Using FVQ to encode the samples in the circle, VQ distortion is reduced. The FVQ encoded circle is closer to the ideal circle than the VQ encoded one is (see Figure 3.1 (c) and (d)). Defining the VQ distortion as the Euclidean distance between the coded form and the input form, results of using FVQ processing to reduce VQ distortion is shown in Figure 3.2. In this two-dimensional example, the distortion of FVQ

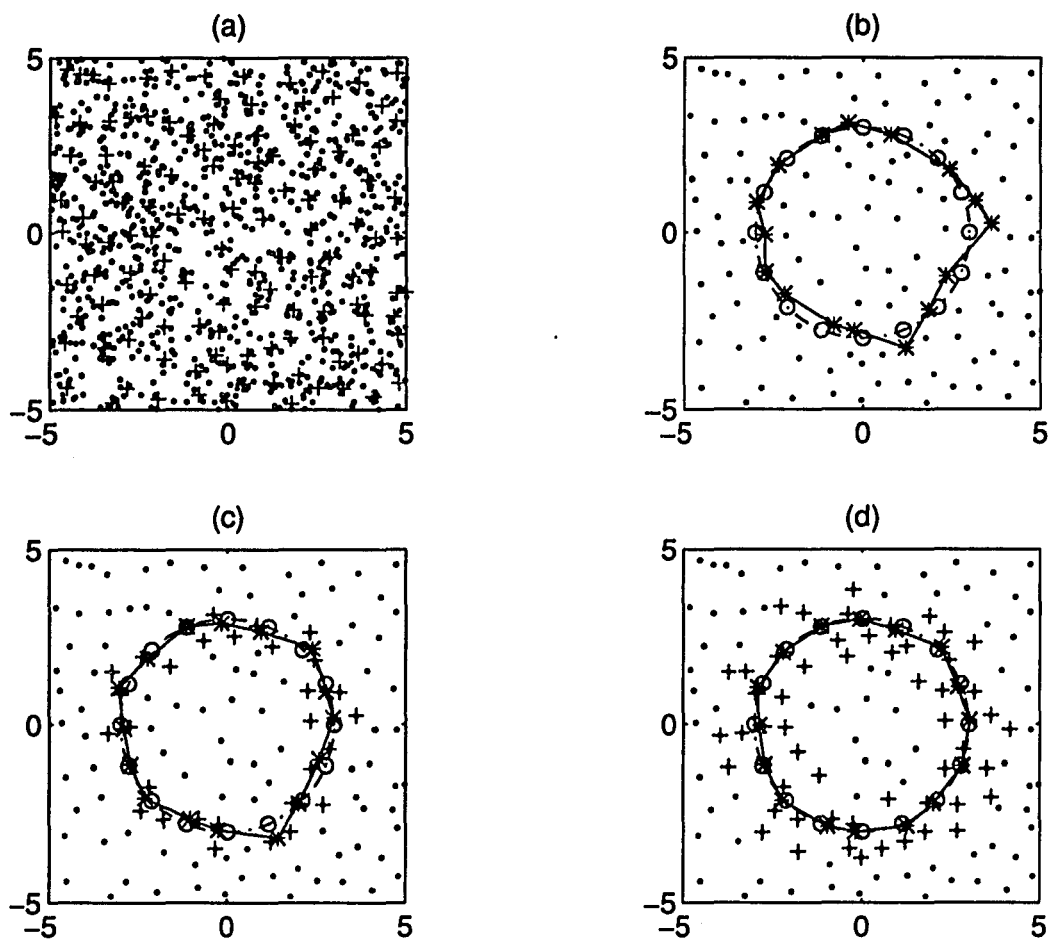


Figure 3.1: Vector quantization and fuzzy vector quantization: (a) samples and their codebook; (b) VQ encoding; (c) FVQ encoding with 2 participating codewords; (d) FVQ encoding with 6 participating codewords. In Figure (a), “.” denotes data samples and “+” denotes codewords. In Figures (b), (c), and (d), “o” denotes input vectors, “*” denotes encoded vectors, “+” denotes involved codewords, and “.” denotes other codewords

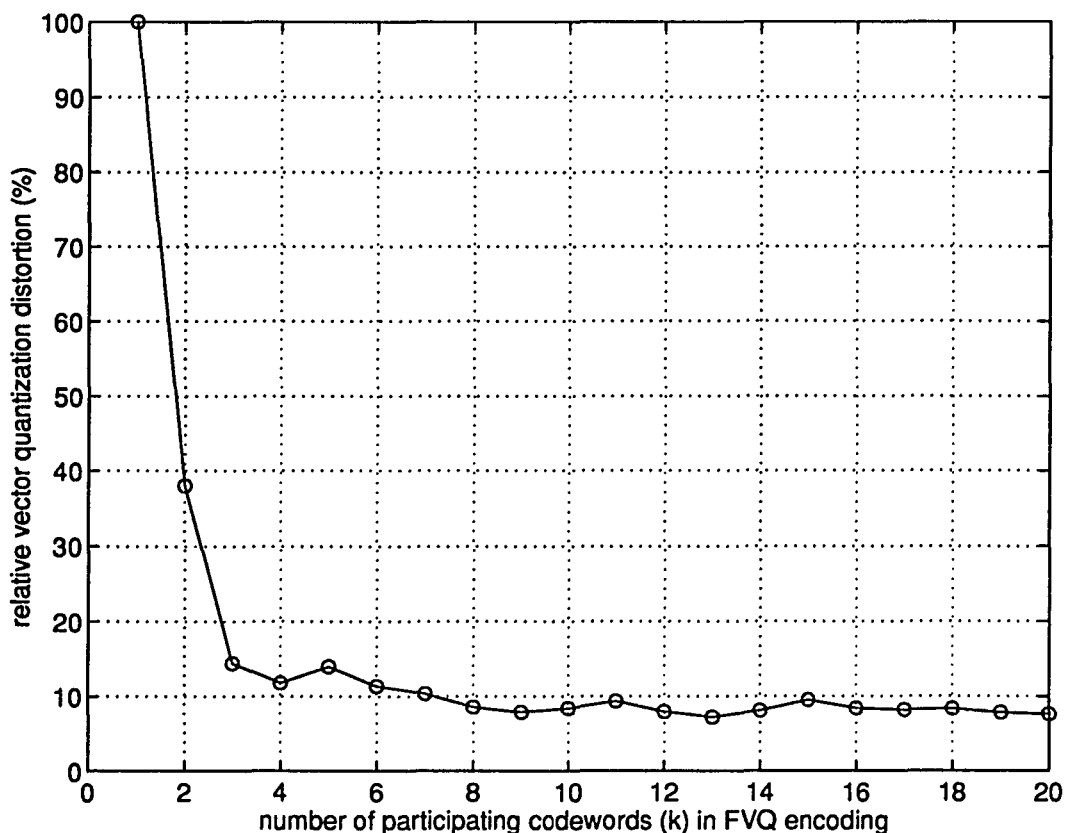


Figure 3.2: Distortion of vector quantization and fuzzy vector quantization: $k=1$ is VQ encoding, and else are FVQ encoding. The VQ distortion is set to 100%

processing is about 10% of VQ processing. A good performance is achieved when $k = 3$. No significant improvement is shown when k increases further.

The vector quantization of speech spectra is equivalent to segment the spectral feature space into multiple different phonemic pieces. This segmentation leads to a codebook with N codewords that are well-distributed in feature space. In principle, a codebook comprising all kinds of phonemes can be used to encode any input utterance. An example of VQ/FVQ encoding of real speech signals is shown in Figure 3.3. The spectra of utterance “sail” shown in Figure 3.3 (a) is VQ-encoded, and

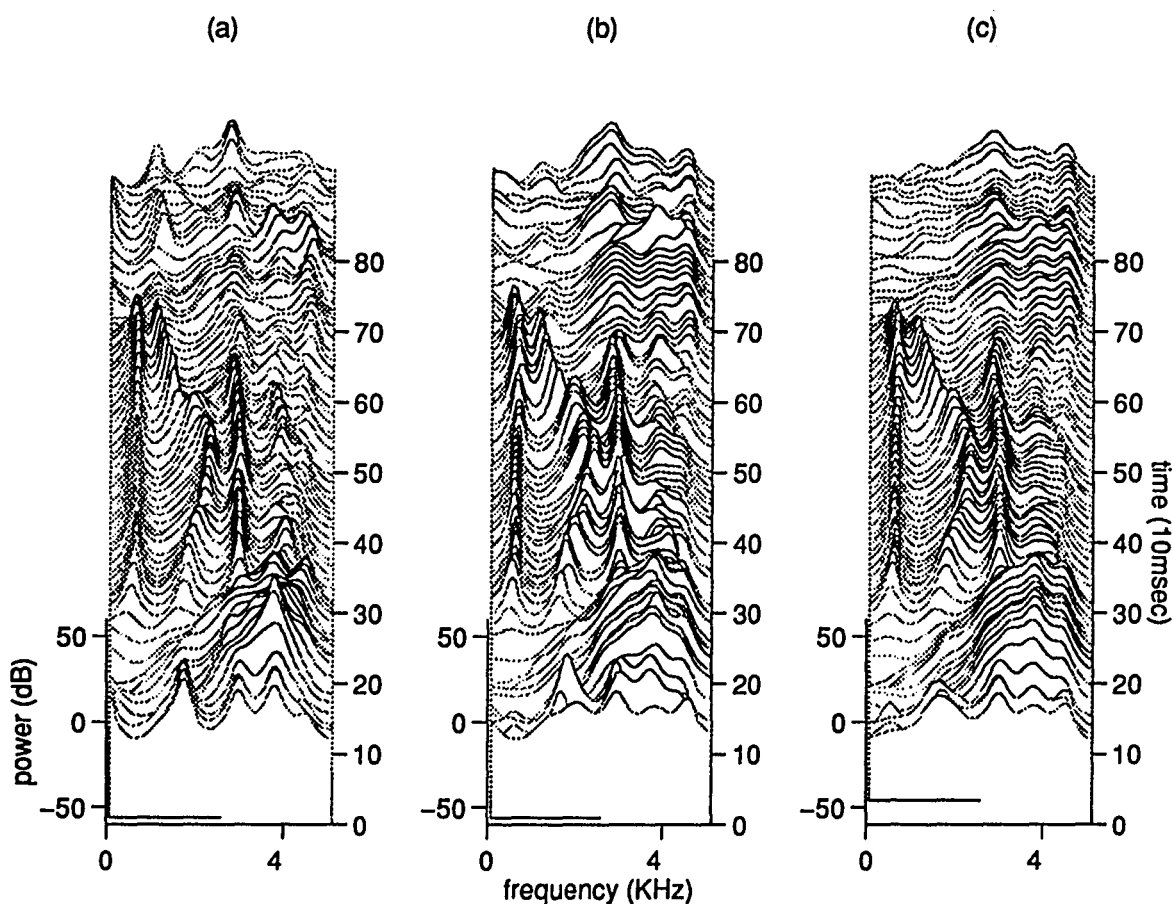


Figure 3.3: Illustration of VQ and FVQ processing by three dimensional spectra of the word "sail": (a) original, (b) VQ encoding, and (c) FVQ encoding

the encoded spectra are shown in Figure 3.3 (b). It can be seen that the formant transition of the encoded speech is not smooth because of VQ distortions. These distortions could cause audible noise in synthetic speech. The FVQ encoded spectra, however, exhibit much smoother formant transition (see Figure 3.3 (c)). Overall, it seems that FVQ could be used to minimize quantization distortions.

3.1.2 Feature Space Mapping and Fuzzy Feature Space Mapping

The mapping information of two feature spaces was obtained through a supervised learning procedure. In this procedure, thousands of the corresponding pairs (one is the vector in the input feature space and the other is the corresponding vector in the target feature space) were collected to construct a correspondence table. We could have use this correspondence table to perform feature space mapping. But, this method is impractical because it requires that all of the training data be loaded and checked.

The advantage of the VQ-based approach is that it permits the feature space mapping to be much less complex. The VQ-based approach needs only to address variations among N vectors, rather than all possible vectors. Once the mapping rules of these N vectors are found, the entire feature space mapping can be accomplished using the input codebook and its corresponding mapping codebook. The mapping codebook was generated using the following steps:

1. All input vectors in the training pairs of the correspondence table were encoded by the input codebook.
2. The target vectors of each input codeword were identified based on the correspondence table.
3. The average of these target vectors was designated to be the target codeword in the mapping codebook.

For comparison, the fuzzy vector quantization (FVQ) was also implemented to compute the mapping codebook. In this approach, each input training vector was encoded by k -nearest codewords. The fuzzy membership was a function that is inversely proportional to the distance between the input vector and these codewords (see equation 3.3). The mapping codeword was computed as a weighted average of all projections from a given input codeword. The weights were equal to the fuzzy memberships.

The final feature space conversion was made by encoding the input vectors using input codebook and decoding these codes using mapping codebook. Suppose \vec{v} is the input vector. $\vec{c}_n (1 \leq n \leq N)$ is the input codebook, and $\vec{m}_n (1 \leq n, m_n \leq N)$ is the mapping codebook. The encoding is to find a code \vec{c}_I which is closest to the input vector \vec{v} . The decoding is to find the target codeword, \vec{m}_I of input codeword \vec{c}_I . The conversion is to replace \vec{v} by \vec{m}_I .

To investigate the properties of feature space mapping, the squared area shown in Figure 3.1 (a) was used as the input feature space. The input codebook was the 128 codewords derived from the 1000 samples. The mapping codebook was generated based on data on the circles and triangles shown in Figure 3.4 (a). The 16 samples on each circle are required to be mapped to the 16 corresponding samples on each triangle. The 32 pairs of vectors form a correspondence table. The mapping relationship, in general, is nonlinear between the feature space of source and target

(where the triangles are located). It is similar to the case of speech spectral conversion where different phonemes should be converted in different ways.

For testing results of feature space mapping, we took a circle as the input data with a radius between those of training circles (see Figure 3.4 (b)). When VQ was adopted to train the system, the mapped form is unacceptable. The main reason is that the mapping relationship of some codewords has not been set up through the training process when the training data is insufficient. If FVQ was adopted in training procedure while using the same training data, the mapped forms became reasonable (see Figure 3.4 (c) and (d)). Better results were achieved because FVQ can supply more training chances to the conversion system than VQ does. This can be depicted in Figure 3.1: if conventional VQ is used to encode the training sample, only the codeword nearest to the sample is involved in the training procedure (see Figure 3.1 (b)). If FVQ was used to encode the sample, k -nearest neighboring codewords are involved in the training procedure (see Figure 3.1 (c) and (d)).

FVQ represents an input vector as a weighted linear combination of codewords with different fuzzy memberships; therefore, a fuzzy mapping can be realized by mapping multiple codewords simultaneously rather than only one. The mapped vector is a linear combination of mapped codewords while preserving the input vector's fuzzy memberships in the source feature space. Formally, by replacing \vec{C}_i in equation 3.2 with the corresponding codeword \vec{M}_i in the mapping codebook while preserving f_i ,

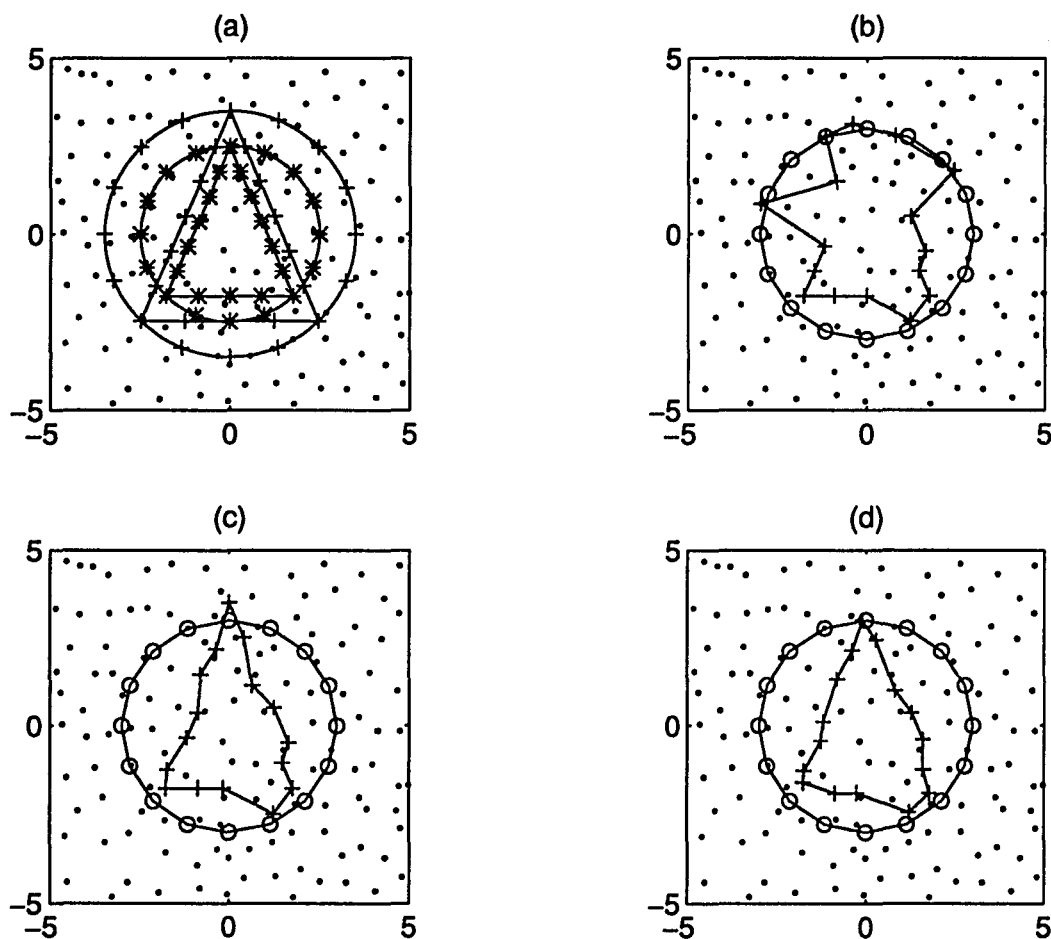


Figure 3.4: VQ-based feature space mapping: (a) training data: two circles are source and two triangles are target; (b) mapping with conventional VQ training; (c) with FVQ training when $k=2$; (d) with FVQ training when $k=6$. In Figure (a): “+” on the big circle and big triangle denotes one group of training pairs, “*” on the small circle and small triangle denotes the other group of training pairs; in Figure (b), (c), and (d): “o” denotes testing input vectors, “+” denotes converted vectors, and “.” denotes codewords in the input codebook

we obtain the mapped vector \vec{t} ,

$$\vec{t} = \sum_{i=1}^k f_i \vec{M}_i. \quad (3.4)$$

The fuzzy mapping procedure leads to a good mapping performance. The mapped form was very similar to the ideal one, not only when the input was a circle (see Figures 3.5 (a)), but also when the input was a rhombus (see Figures 3.5 (c)). The performance of fuzzy mapping with insufficient training data was very close to the performance of that with sufficient training data (see Figures 3.5 (b) and (d)), where training data, comprising 100 circles with different radius were taught to convert to 100 corresponding triangles. It is worth mentioning that the mapped forms in Figures 3.5 (b) and (d) do not achieve 100% perfection, even based on using sufficient training data.

Defining the mapping error or distortion as the Euclidean distance between the mapped form and the ideal target form, the effective of fuzzy mapping with different k is shown in Figure 3.6. The best result is achieved when $k = 4$. The mapping distortion of fuzzy mapping is only 27% of the conventional mapping. No more improvement is shown when k increases further.

Our results show that the FVQ and fuzzy mapping can be used to enhance the system performance and overcome the problem of insufficient training data; however, the VQ distortion cannot be completely avoided in the VQ-based feature space mapping system.

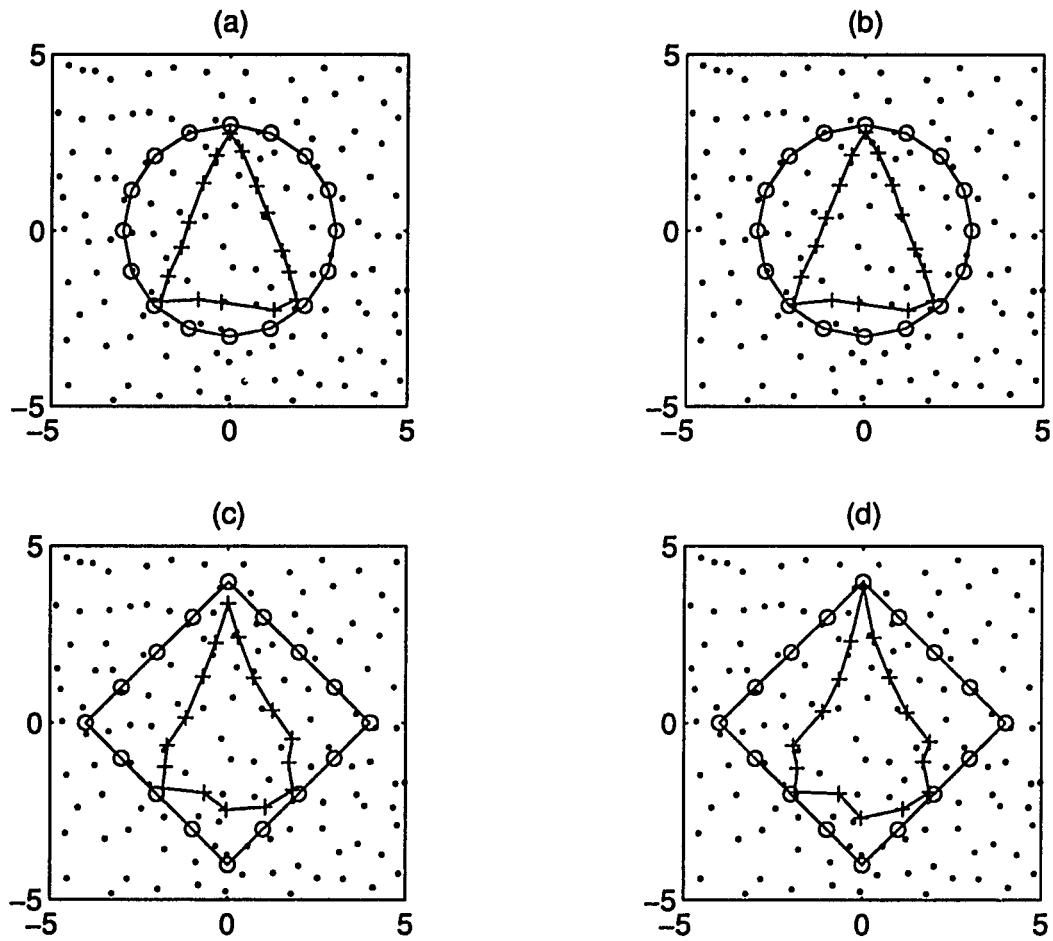


Figure 3.5: VQ-based feature space fuzzy mapping: (a) mapping of testing circle and (c) testing rhombus with insufficient training; (b) and (d) with sufficient training. “o” denotes testing input vectors, “+” denotes converted vectors, and “.” denotes codewords in the input codebook

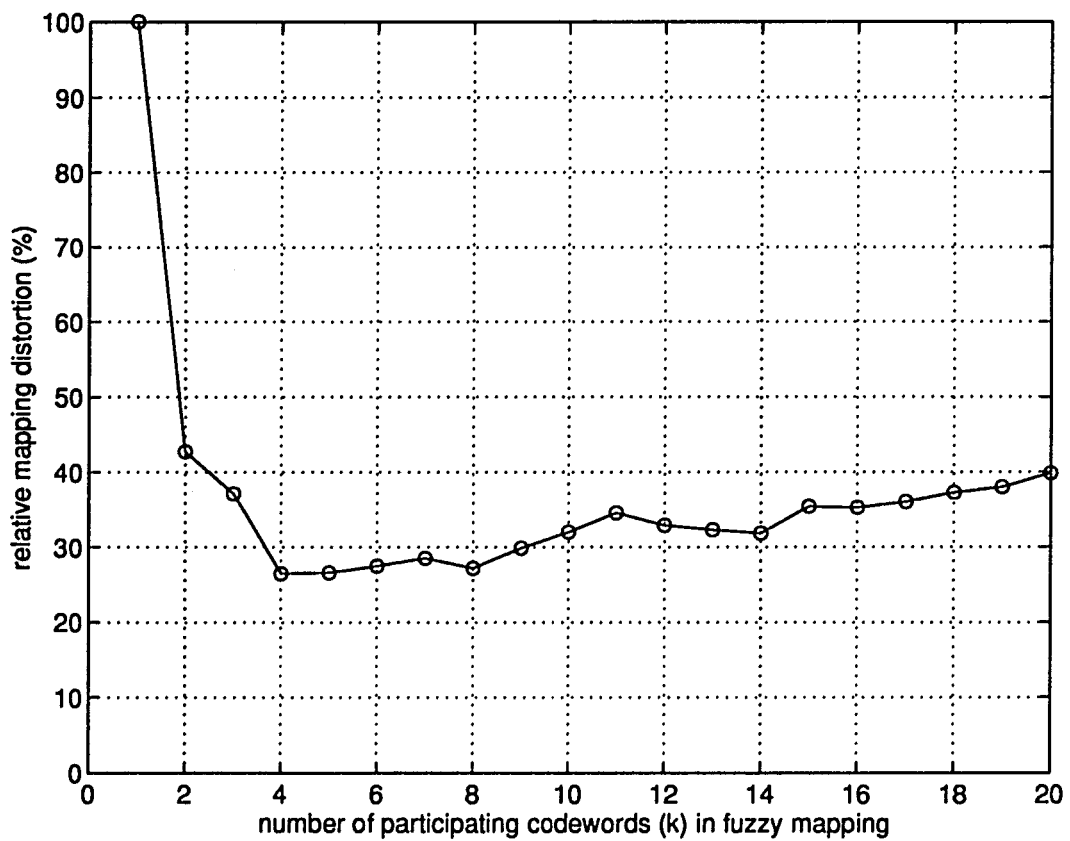


Figure 3.6: Distortion of fuzzy mapping: $k=1$ is conventional mapping, and else are fuzzy mapping. The conventional mapping distortion is set to 100%

The advantage of using FVQ to train the mapping codebook was also revealed by the distribution of the training vector number in each codeword (see Figure 3.7). If the VQ was used, there were 40 codewords whose membership of mapping training vector-pairs is less than 5, which might be insufficient to generate a suitable codeword in the mapping codebook. If the FVQ was used, there were only 10 codewords in the input codebook whose accumulated membership of training vector-pairs was less than 5. The histogram of its membership was close to a normal distribution. Therefore, given the same training data, the information will be used more effectively to build the mapping relationship if the FVQ process is adopted.

One example of VQ-based spectral conversion is shown in Figure 3.8. A male alaryngeal speaker and a male normal speaker joined the spectral conversion system training. FVQ and fuzzy mapping were used in the learning and conversion phases. In this testing example, voiced segmentation of the utterance "sail" by alaryngeal speech was converted to that of normal speech. It is clear, from the three-dimensional spectra, that the formants of vowel in synthetic speech are lower than those in alaryngeal speech. The synthetic speech sounds like the target speech.

Our testing results show that the converted speech usually sounds smooth but a little ambiguous. The reason is that the minute spectral temporal turbulence in an utterance is dismissed by the VQ process and mapped spectra have a wider formant bandwidth; therefore, the formant enhancement technique was studied in this work to improve the performance of VQ-based speech conversion (see next chapter).

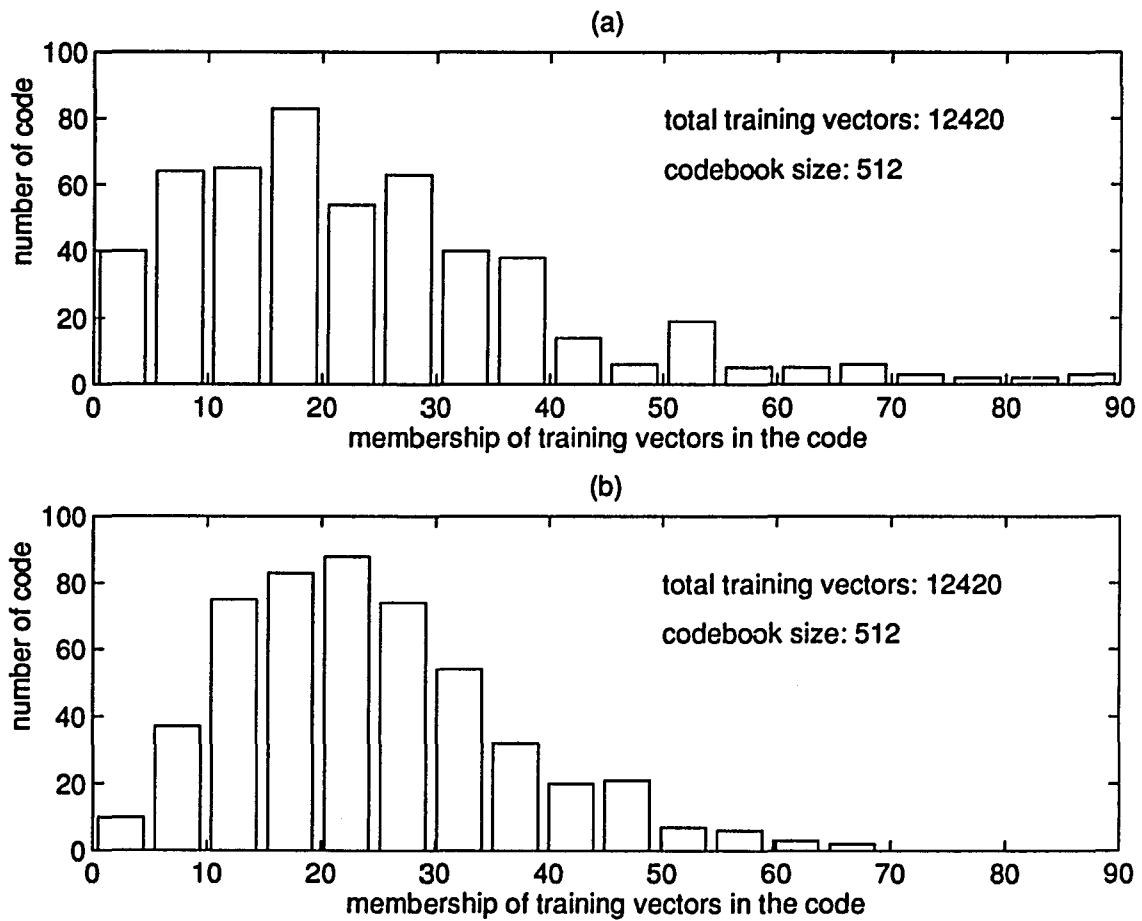


Figure 3.7: Distribution of the number of the training vector-pairs in each codeword: (a) using the VQ process, and (b) using the FVQ process

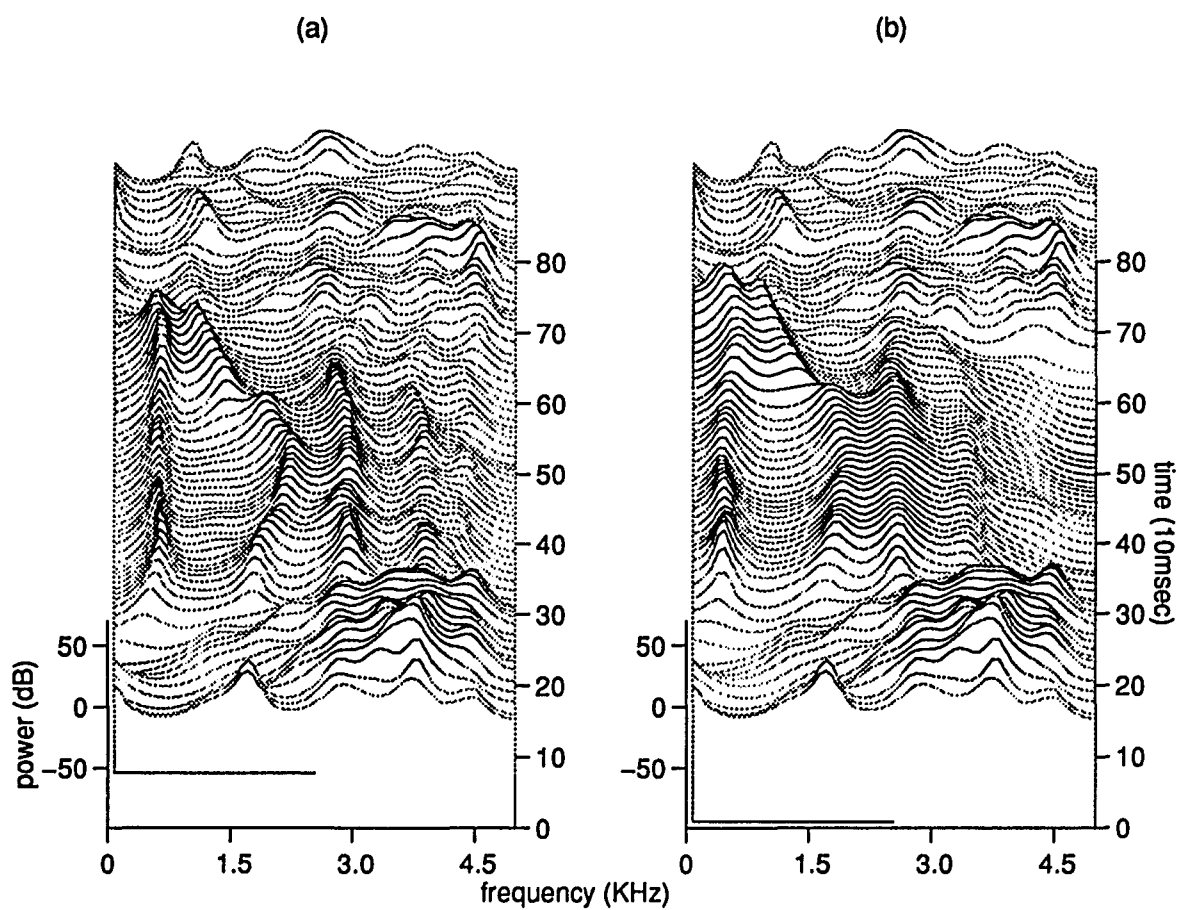


Figure 3.8: Illustration of VQ-based speech conversion by three dimensional spectra of the word "sail": (a) alaryngeal speech; (b) converted speech by VQ-based approach

3.2 LMR-Based Spectral Conversion

From the correspondence table generated in the supervised learning procedure, it is clear that spectral conversion can be thought as a paired-associate problem, i.e., for a given spectrum of input speech, we are asked to find the corresponding spectrum of target speech. If spectra can be assumed to be transformed linearly, we may treat it as an optimal linear associative mapping problem that can be solved by linear multivariate regression (LMR).

3.2.1 Linear Multivariate Regression

Assuming that an input vector \vec{x}_k belongs to the source feature space, an output vector \vec{y}_k belongs to the target feature space, and both of them are p -dimensional vectors (it is true in our case since both \vec{x}_k and \vec{y}_k are LPC cepstral coefficients with the same order), a linear associative mapping can be defined as

$$\vec{y}_k = M\vec{x}_k, \quad k = 1, 2, \dots, m, \quad (3.5)$$

where M is a $p \times p$ matrix used to keep the mapping information between source and target spaces. We may regard \vec{y}_k as the memorized data and \vec{x}_k as the search argument by which \vec{y}_k is encoded and retrieved (Kohonen, 1977; Kohonen, 1989).

There is a solution to the pair-associate problem that is optimal in the sense of least squares. If there exists m pairs of training vectors, we can define matrices Y

and X with \vec{y}_k and \vec{x}_k as their columns, respectively, i.e.,

$$Y = [\vec{y}_1, \vec{y}_2, \dots, \vec{y}_m], \quad X = [\vec{x}_1, \vec{x}_2, \dots, \vec{x}_m]. \quad (3.6)$$

Then, a matrix form of equation 3.5 is

$$Y = MX, \quad (3.7)$$

or

$$[\vec{y}_1, \vec{y}_2, \dots, \vec{y}_m] = M[\vec{x}_1, \vec{x}_2, \dots, \vec{x}_m]. \quad (3.8)$$

To minimize the mean-square error between the target vectors and the retrieved vectors from the linear associative mapping, i.e.,

$$\sum_{k=1}^m \|\vec{y}_k - M\vec{x}_k\|^2, \quad (3.9)$$

the matrix M can be obtained by multiplying Y by the pseudo-inverse of X :

$$M = YX^\dagger, \quad (3.10)$$

or

$$M = [\vec{y}_1, \vec{y}_2, \dots, \vec{y}_m][\vec{x}_1, \vec{x}_2, \dots, \vec{x}_m]^\dagger, \quad (3.11)$$

where † denotes the pseudo-inverse. The pseudo-inverse of X can be calculated by

$$X^\dagger = X^T(XX^T)^{-1}, \quad (3.12)$$

where T denotes the matrix transpose, and $^{-1}$ denotes the matrix inverse. The method of solving this least squares problem is called the linear multivariate regression (LMR). The solution of M is optimal in the sense of least squares to a set of

given training pairs. In such a system, any input vector that resembles any of the stored training vectors is transformed approximately to the vector in the associated training pair.

Obviously, the capacity of the LMR-based system is limited by the dimension of matrix M . For example, by using the LMR to deal with the problem of two-dimensional feature space mapping mentioned before, the retrieved form of an input circle was an ellipse rather than a triangle (see Figure 3.9 (a)), where the training data consist of 100 circles and corresponding triangles. This result is predictable because the LMR-based system merely has the ability to perform a linear transformation. For a two-dimensional space, the linear transformation only includes scale changing, original point shifting, and rotation. The limited capacity results in a problem when the LMR was used to implement speech spectral conversion. Indeed, it is difficult to imagine that varied vowels and consonants can share similar spectral transformation properties; therefore, the multi-subset approach has to be used.

3.2.2 Capacity of Multi-Subset LMR

Because a single mapping matrix cannot accommodate diversified mapping relationships, a multi-subset approach was used in the speech conversion (Valbret et al., 1992). In this method, the source feature space is segmented into multi-subsets to decrease the mapping complexity. Applying LMR to each subset, a nonlinear mapping

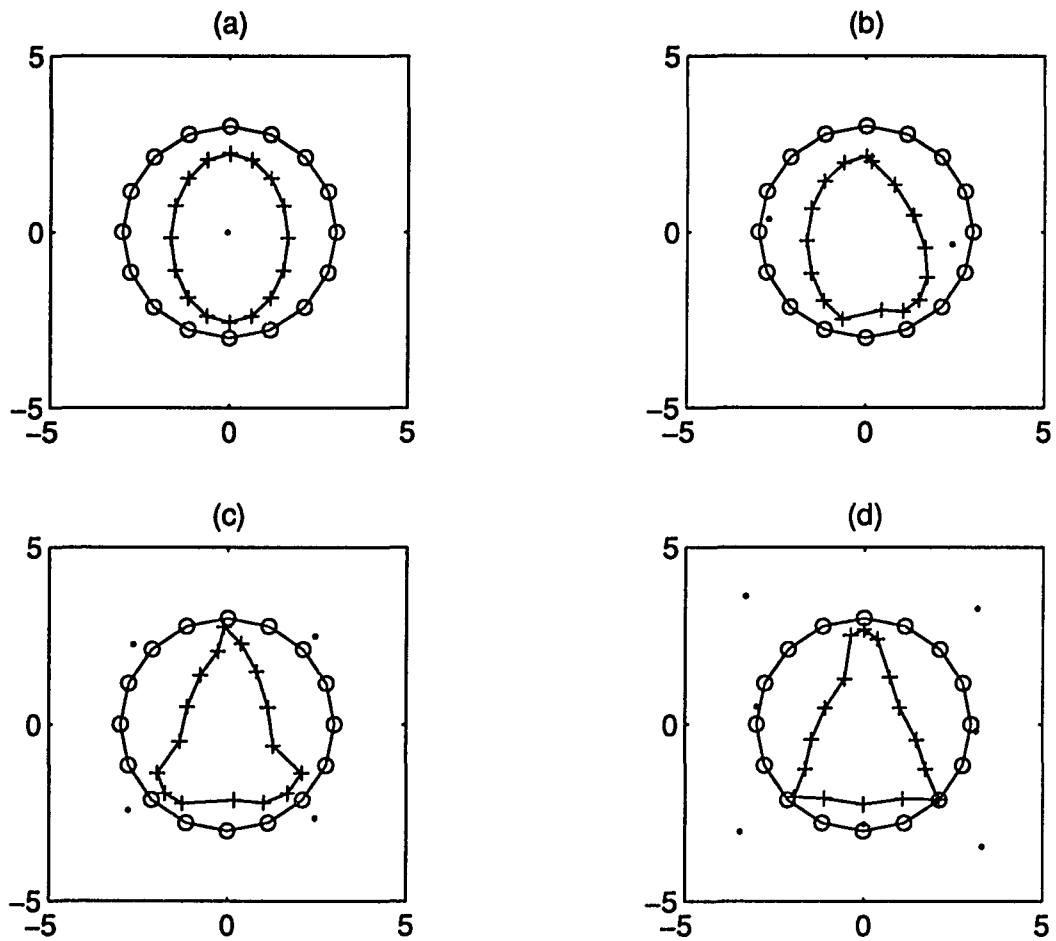


Figure 3.9: LMR-based feature space mapping: (a) one-subset, (b) two-subsets, (c) four-subsets, and (d) eight-subsets. “o” denotes testing input vectors, “+” denotes converted vectors, and “.” denotes the center of subsets

is realized through a combination of multi-linear mapping. The subsets are generated by clustering the source feature space using VQ technique.

Assuming that the set of source vectors in the training pairs belonging to the n th subset is denoted by $\bar{v}_j^{n,s}$, where $j = 1, 2, \dots, m_n$; and the set of corresponding target vectors is denoted by $\bar{v}_j^{n,t}$, where $j = 1, 2, \dots, m_n$, means of this subset for the source and target vectors are computed, respectively:

$$\bar{v}^{n,s} = \frac{1}{m_n} \sum_{j=1}^{m_n} \bar{v}_j^{n,s}, \quad \bar{v}^{n,t} = \frac{1}{m_n} \sum_{j=1}^{m_n} \bar{v}_j^{n,t}. \quad (3.13)$$

Normalization of training vectors are preferred

$$\dot{\bar{v}}_j^{n,s} = \bar{v}_j^{n,s} - \bar{v}^{n,s}, \quad \dot{\bar{v}}_j^{n,t} = \bar{v}_j^{n,t} - \bar{v}^{n,t}, \quad j = 1, 2, \dots, m_n. \quad (3.14)$$

The mapping matrix of the n th subset is

$$M^n = [\dot{\bar{v}}_1^{n,t}, \dot{\bar{v}}_2^{n,t}, \dots, \dot{\bar{v}}_{m_n}^{n,t}] [\dot{\bar{v}}_1^{n,s}, \dot{\bar{v}}_2^{n,s}, \dots, \dot{\bar{v}}_{m_n}^{n,s}]^\dagger. \quad (3.15)$$

Testing results of the two-dimensional space mapping problem are shown in Figure 3.9 (b), (c) and (d), where the number of subset is 2, 4, and 8, respectively. The mapping distortion under different number of subset is shown in Figure 3.10. When the number of subsets increased, better mapping results are achieved. The selection of subset number is related to the complexity of mapping task. In principle, we hope to find the minimum number of subsets to accomplish the given task because less subsets require less data space.

It is reasonable to expect the LMR-based approach can achieve better performance than the VQ-based approach because the mapping in any given region of the feature

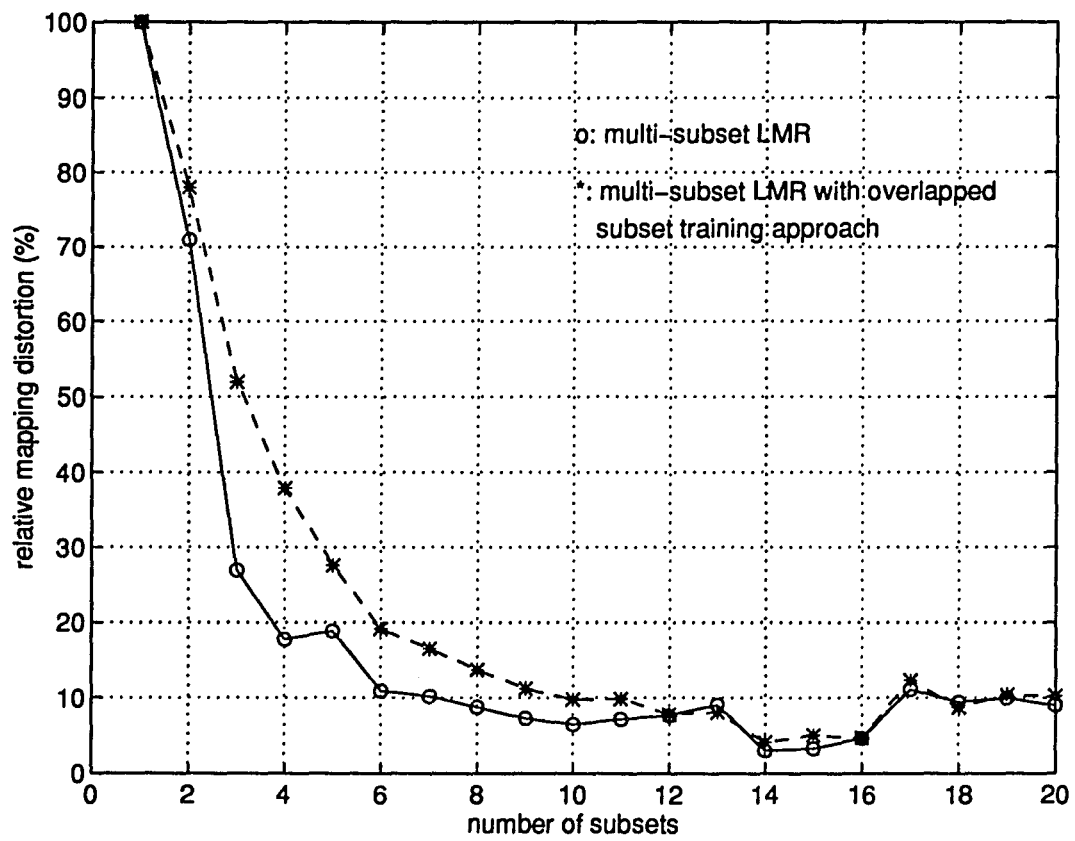


Figure 3.10: Distortion of multi-subset LMR mapping

space is continuous; thus, the distortion due to quantization noise and feature-vector averaging were essentially eliminated. Using the two-dimensional feature space mapping problem as an example, when there are sufficient training data, the performances of multi-subset LMR are perfect (see Figure 3.11 (b) and (d), where the number of subset is 16 and the training data are 100 circles and triangles). Comparing it with the performance of VQ-based approach shown in Figure 3.5 (b) and (d), the outputs of multi-subset LMR-based mapping system have less distortion.

Implementing the LMR-based system, we noticed that synthetic speech is more flexible to the input utterance than the output of the VQ-based system. The problem with this method is that synthetic speech sometimes displays audible distortions that have not as yet been fixed (Valbret et al., 1992), a problem that decreases its application value.

Based on the discussions and evaluations presented above, it seems that the VQ-based conversion is more desirable when the size of the learning data set is limited. The major problem of VQ-based approach is that the quantization distortion and distortions due to vector-averaging can be problematic. The LMR-based spectral conversion seems to be able to avoid the problems associated with the VQ-based approach; however, discontinuities during transitions between subsets may cause some problems. These problems need to be addressed before either the VQ-based or LMR-based conversion algorithm can be used for speech enhancement.

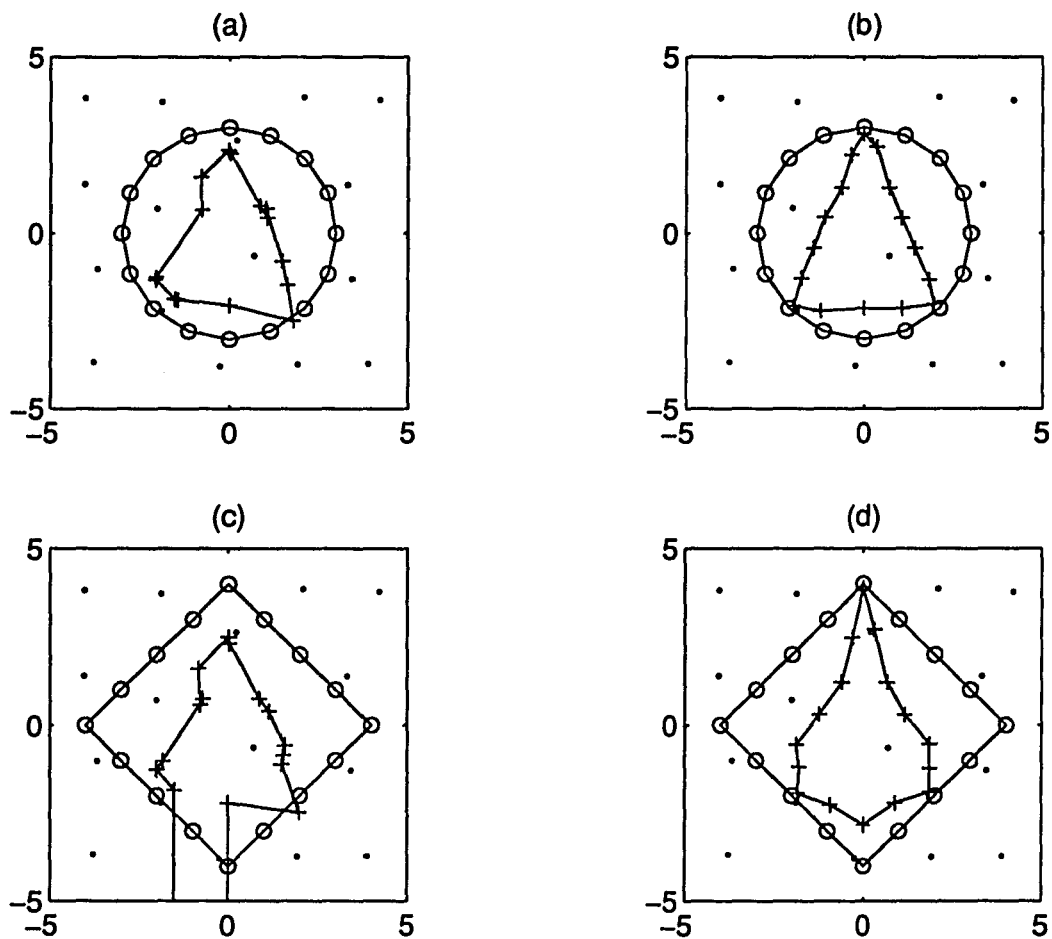


Figure 3.11: Multi-subset LMR with and without sufficient training: (a) and (c) with insufficient training data; (b) and (d) with sufficient training data. "o" denotes testing input vectors, "+" denotes converted vectors, and "." denotes the center of subsets

CHAPTER 4

MODIFICATIONS OF SPEECH CONVERSION METHODS

In this chapter, modifications of the original VQ- and LMR-based spectral conversion methods are presented. These modifications are aimed at reducing the spectral distortion (bandwidth increase) in the VQ-based method and the spectral discontinuity in the LMR-based method.

4.1 Modification of VQ-Based Conversion Method

The bandwidth increase in the VQ-based speech conversion system is intrinsic to the algorithm of vector quantization. Vector quantization is an algorithm for choosing a set of codewords (spectra) that *optimally* represent the whole spectral space of a given speaker. Each codeword eventually is an average of a small cluster of spectra because average is the *optimal* representation of spectra in its vicinity. Unfortunately, the average spectrum also tends to have a larger bandwidth than its constituents.

The bandwidth increase is also intrinsic to the VQ-based conversion mapping scheme, where the target spectrum is designated as the average of all the spectra projected from a given cluster in the input spectral space. A small cluster in the input spectral space might project divergently to a large area in the target spectral

space. When the divergent projection occurs, the bandwidth of the target spectrum will be large.

Perceptually, speech synthesized with large bandwidths sounds ambiguous and unclear. Because spectral averaging cannot be avoided in the VQ-based spectral conversion system, formant enhancement in the speech conversion process was included to compensate for the bandwidth increase. Formant enhancement was made after spectral conversion and before speech synthesis.

4.1.1 Formant Enhancement Using Chirp Z-Transform

One method to sharpen the spectral peaks/formants is to use the chirp z-transform (Rabiner et al., 1969). The chirp z-transform allows for the evaluation of a transfer function on a contour that is not the unit circle. If the contour for computing spectral transfer function is located outside all poles of the transfer function and inside the unit circle, the bandwidth of the resulting spectral transfer function will be reduced.

Mathematically, the z-transform of any sequence x_n is defined as

$$X(z) = \sum_{n=-\infty}^{\infty} x_n z^{-n}. \quad (4.1)$$

When $z = e^{j\omega}$, equation 4.1 provides the discrete Fourier transform of x_n :

$$X(j\omega) = \sum_{n=-\infty}^{\infty} x_n e^{-jn\omega}. \quad (4.2)$$

When $z = re^{j\omega}$, where r is an arbitrary complex numbers, equation 4.1 defines the chirp z-transform:

$$X(j\omega) = \sum_{n=-\infty}^{\infty} x_n r^{-n} e^{-jn\omega}. \quad (4.3)$$

A special case of the chirp z-transform is when r is a constant and $|r| < 1$. It yields the z-transform of x_n on a circle with a radius $|r| < 1$.

There are several ways to implement the chirp z-transform. One method is to multiply the LPC coefficients, a_i , by a factor, $a'_i = r^{-i}a_i$, and evaluate the adjusted polynomial on the unit circle (McCandless, 1974),

$$A(j\omega) = 1 + \sum_{i=1}^M a_i r^{-i} e^{-ji\omega} = 1 + \sum_{i=1}^M a'_i e^{-ji\omega}, \quad (4.4)$$

where $A(j\omega)$ is the denominator of the all-pole model (see Equation 2.6). The resulting spectrum will have sharper spectral peaks/formants than the original spectrum because the poles are effectively pushed out toward the unit circle. The magnitude of r , however, is difficult to control. If r is too large, it tends not to have any positive effect. If r is too small, it may make the LPC filter unstable.

An alternative is to implement the chirp z-transformation in the time-domain. By substituting the system impulse response, h_n , with a weighted sequence, $r^{-n}h_n$, the transfer function of the system is evaluated on a circle inside the unit circle. Because the final LPC synthesis filter is obtained from the impulse response using the auto-correlation method, the filter will be stable as long as the modified sequence, $r^{-n}h_n$, is approximately stationary.

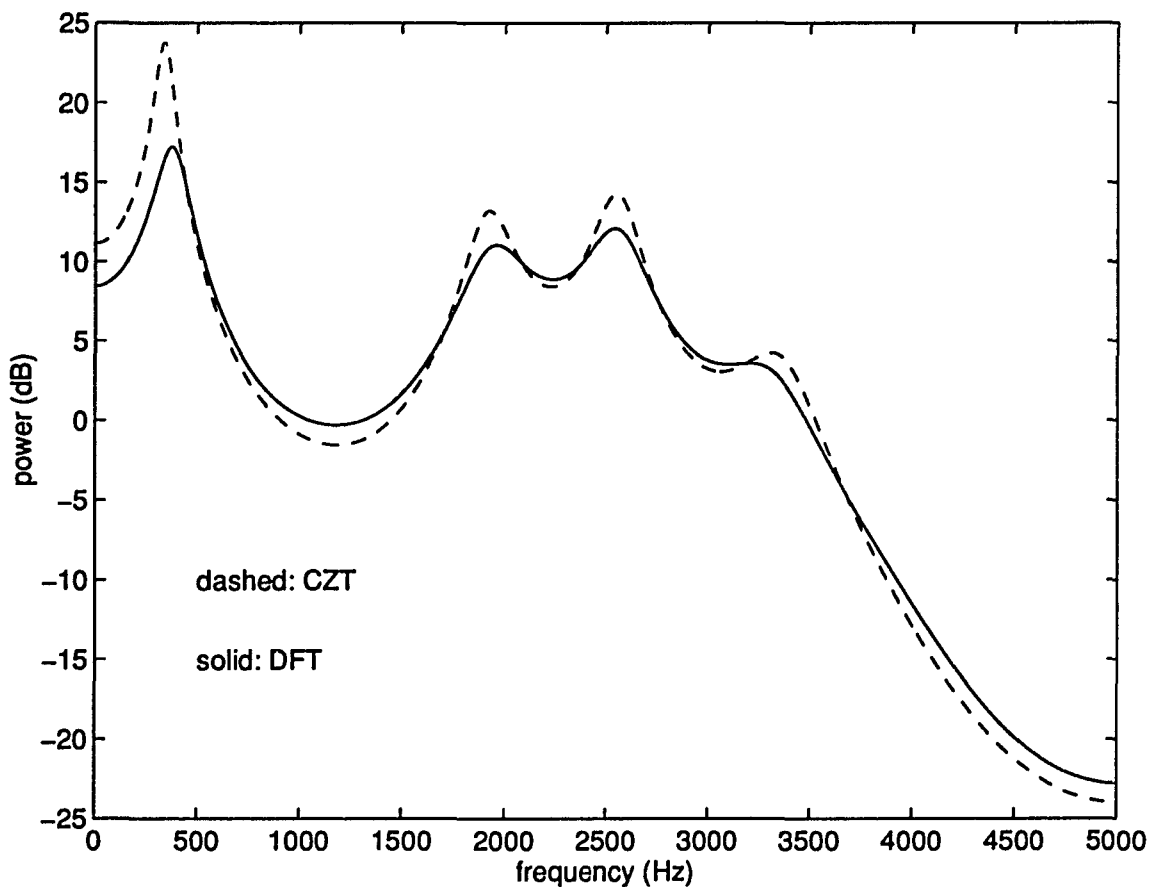


Figure 4.1: Example of formant enhancement using the chirp z-transform

In our VQ-based conversion system, the chirp z-transform was implemented using the weighted impulse response. The impulse response was the minimum phase sequence derived from the inverse process of homomorphic deconvolution. An example of the converted spectrum before and after formant enhancement is shown in Figure 4.1, where $r = 0.96$,

4.1.2 Formants Enhancement Using Cepstral Weighting

The magnitude of formant enhancement using the chirp-z transform is limited by the magnitude of r . To enhance the formants further, the cepstral weighting method (Rabiner and Juang, 1993) in the VQ-based conversion system was applied.

As reviewed in Chapter 2, the cepstrum for the vocal transfer function is a windowed segment of the whole cepstrum. This windowing operation is equivalent to a convolution in the frequency domain between the logarithmic spectrum of the original signal and the spectrum of the rectangular window. The spectrum of the rectangular window is characterized by a narrow mainlobe, but large sidelobes (Oppenheim and Schaffer, 1989). These sidelobes (see Figure 4.2) tend to smooth out the resulting spectrum.

To enhance formants further, the rectangular window was replaced by a more rounded sine window,

$$w(n) = \begin{cases} 1 + h \sin[(n-1) * \pi / (L-1)] & \text{for } n = 1, 2, \dots, L \\ 0 & \text{otherwise} \end{cases} \quad (4.5)$$

where h is a gain factor and is set to less 1. Because the sine window has smaller sidelobes than the rectangular window, it can reduce spectral smoothing to a certain extent. An example of formant enhancement using the sine window is shown in Figure 4.3. An example of formant enhancement by applying both chirp z-transform and sine window is illustrated in Figure 4.4.

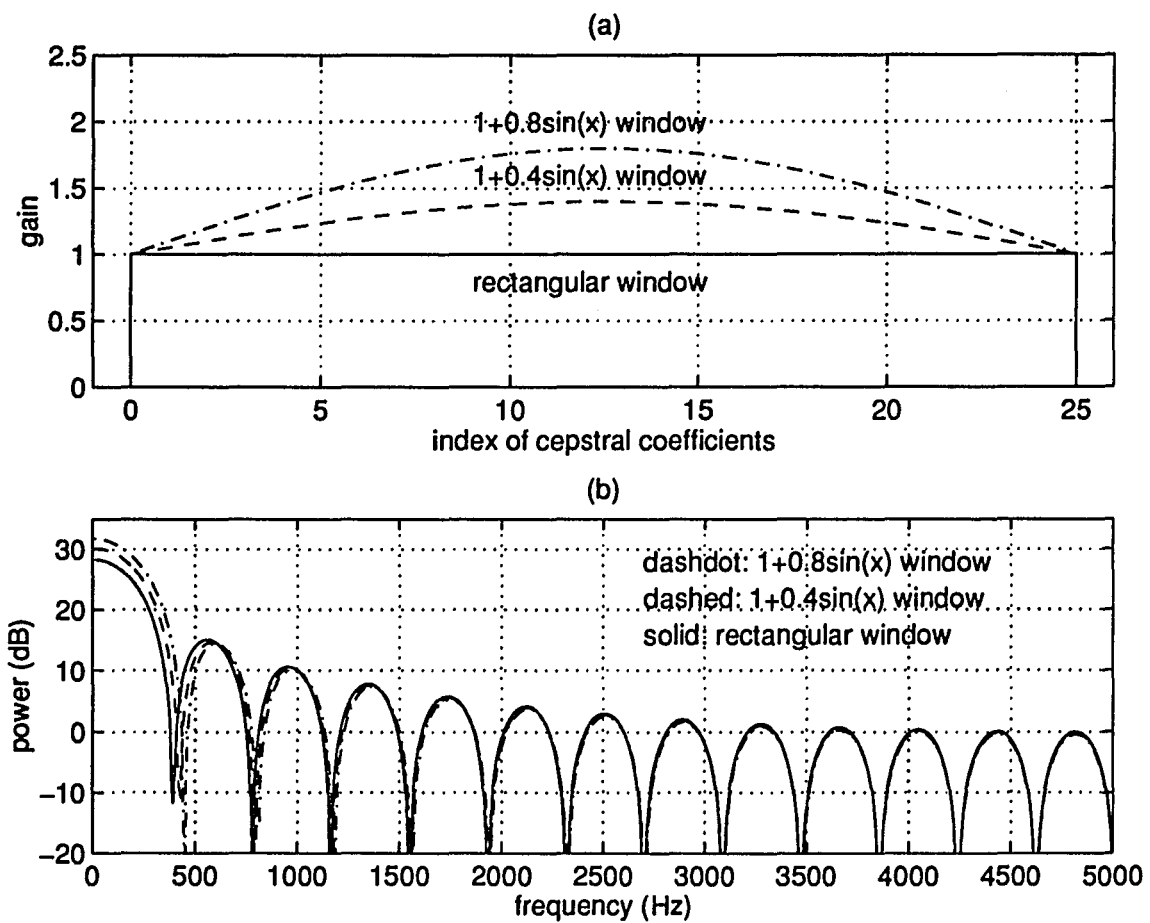


Figure 4.2: Rectangular and sine windows and their corresponding spectra

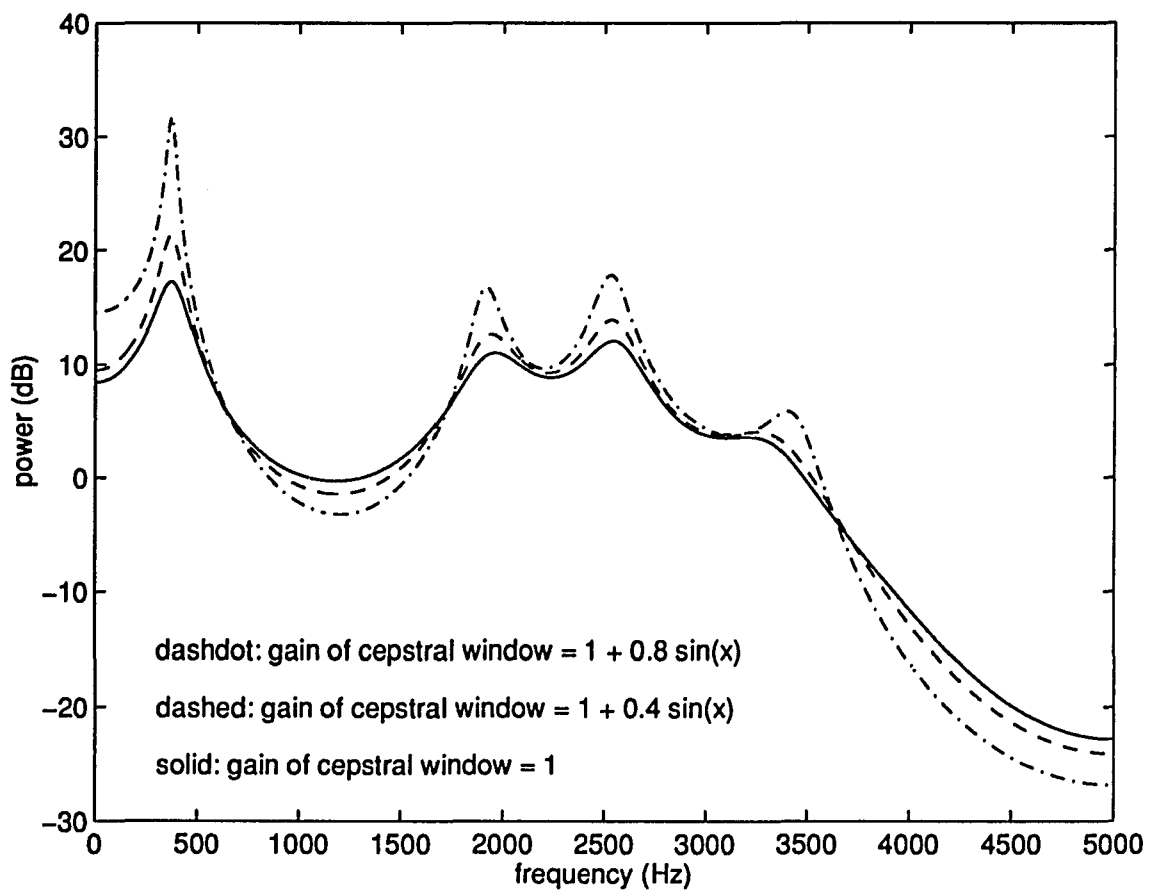


Figure 4.3: Example of formant enhancement using cepstral weighting

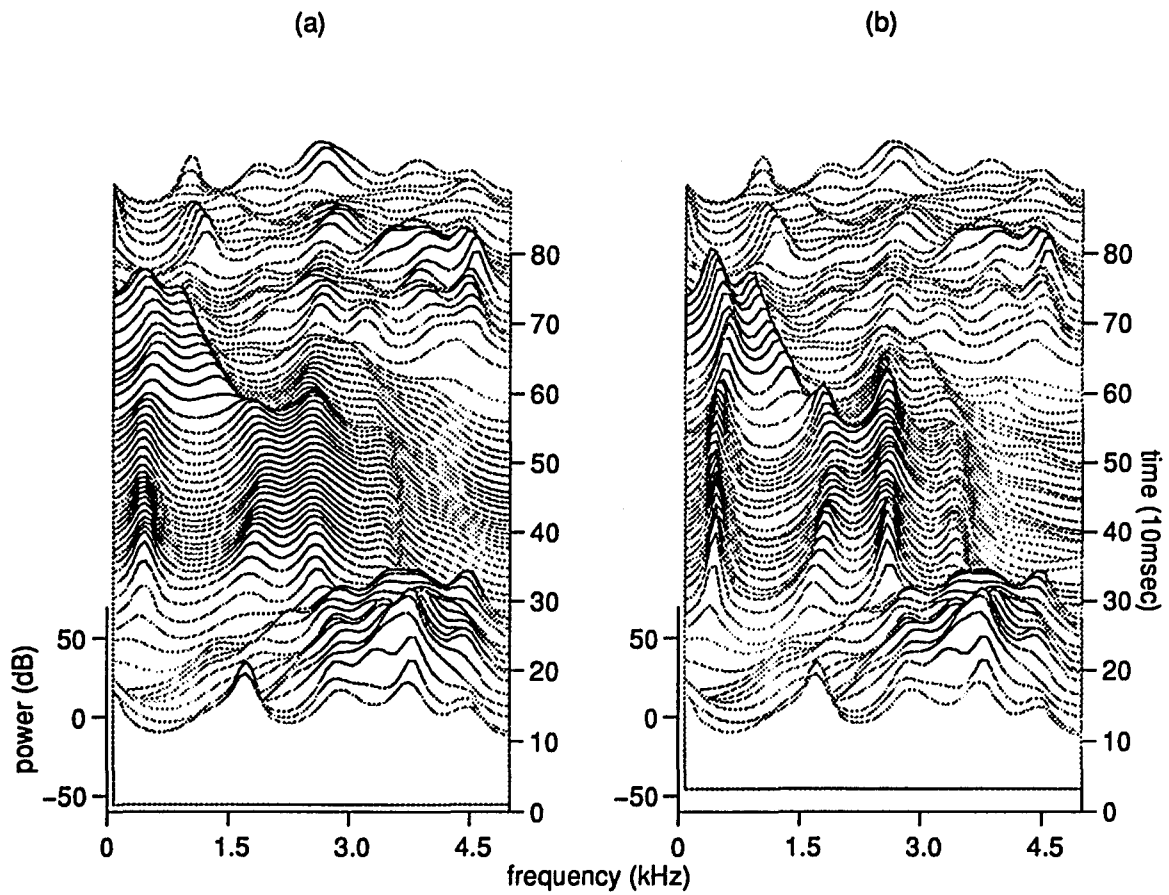


Figure 4.4: Illustration of formant enhancement by three dimensional spectra of the word "sail": (a) after the conventional VQ-based conversion; (b) the modified (formant enhanced) VQ-based conversion

4.2 Modification of LMR-Based Method

In the LMR-based approach, the spectral space was partitioned by a few large clusters and the spectrum within each cluster was mapped continuously (Valbret et al., 1992). The transitions between clusters, however, can be discontinuous resulting in audible clicks in the converted speech output.

This discontinuity is caused by using a non-overlapped subset to derive the LMR mapping matrix. As a result, each mapping matrix is constrained only by samples of a given subset and ignores the behavior of neighbor subsets. While each mapping matrix might serve its constituents satisfactorily, neighboring mapping matrices may target toward different directions, resulting in spectral discontinuities during transitions.

In addition, some subsets may only have a limited number of samples. When the sample size is small, the mapping matrix is constructed from an undetermined rather than an over-determined LMR problem. The solution (mapping matrix) of an undetermined LMR problem can be unpredictable.

4.2.1 Overlapped Multi-Subset Training Approach

To reduce the spectral discontinuity, an overlapped training method is proposed. In this method, overlapped subsets are used to obtain the LMR mapping matrix. A graphic illustration of overlapped training is shown in Figure 4.5.

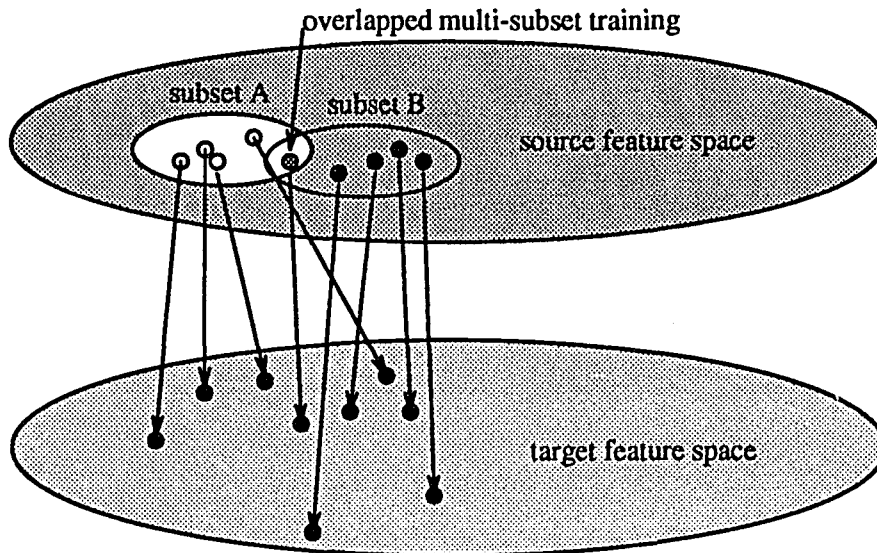


Figure 4.5: Illustration of overlapped multi-subset training approach for LMR-based method

The advantages of using overlapped subsets during training are

- the mapping matrix of each subset is constrained, to a certain extent, by samples of neighboring subsets so that continuity between transitions can be maintained;
- the size of training samples of each subset is effectively increased so that the LMR mapping is likely to be an over-determined problem as it should be.

The membership of a training sample, \vec{v} , is determined by the distance, d_i , between the sample and the cluster centers, $\vec{c}_i, i \in 1, 2, \dots, N$, where N is the total number of cluster. Let $d_1 \leq d_2 \leq \dots \leq d_i \leq \dots \leq d_N$ denote these distances, the training

sample, \vec{v} , will participate in the training of cluster i if

$$D = \frac{d_1}{d_i} \quad (4.6)$$

is greater than a given threshold. The number of clusters that a training sample can join is limited to a maximum I . The overlapped area among neighboring subsets is controlled by the threshold. Obviously, when the threshold is 1, there will be no overlap.

To solve the two-dimensional problem, the new approach avoids the over-shoot in Figure 3.11 (c) when there is no sufficient training data (see Figure 4.6 (a) and (c)). It also performs perfectly when there is sufficient training data (see Figure 4.6 (b) and (d), and Figure 3.10).

An example of using overlapped training in LMR-based spectral conversion is shown in Figure 4.7, where the threshold is 0.75.

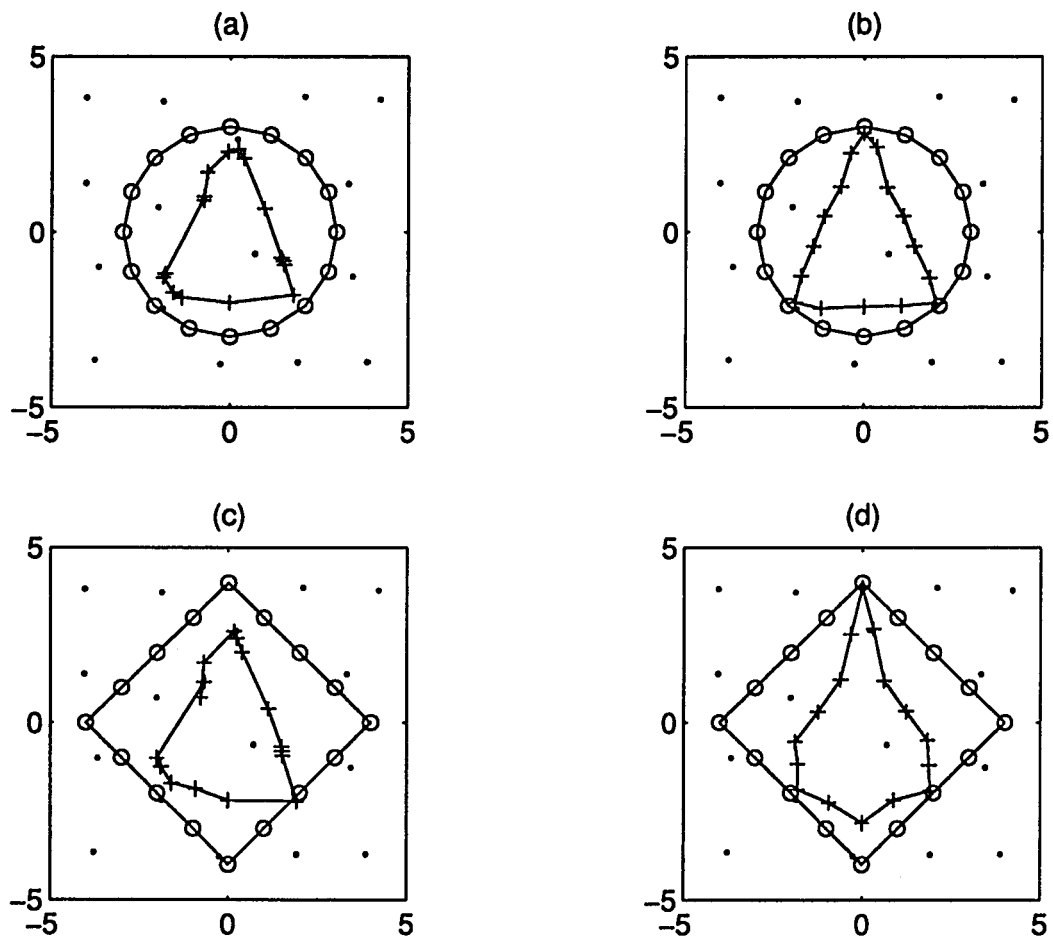


Figure 4.6: LMR-based feature space mapping with overlapped subset training approach: (a) and (c) with insufficient training data; (b) and (d) with sufficient training data. “o” denotes testing input vectors, “+” denotes converted vectors, and “.” denotes the center of subsets

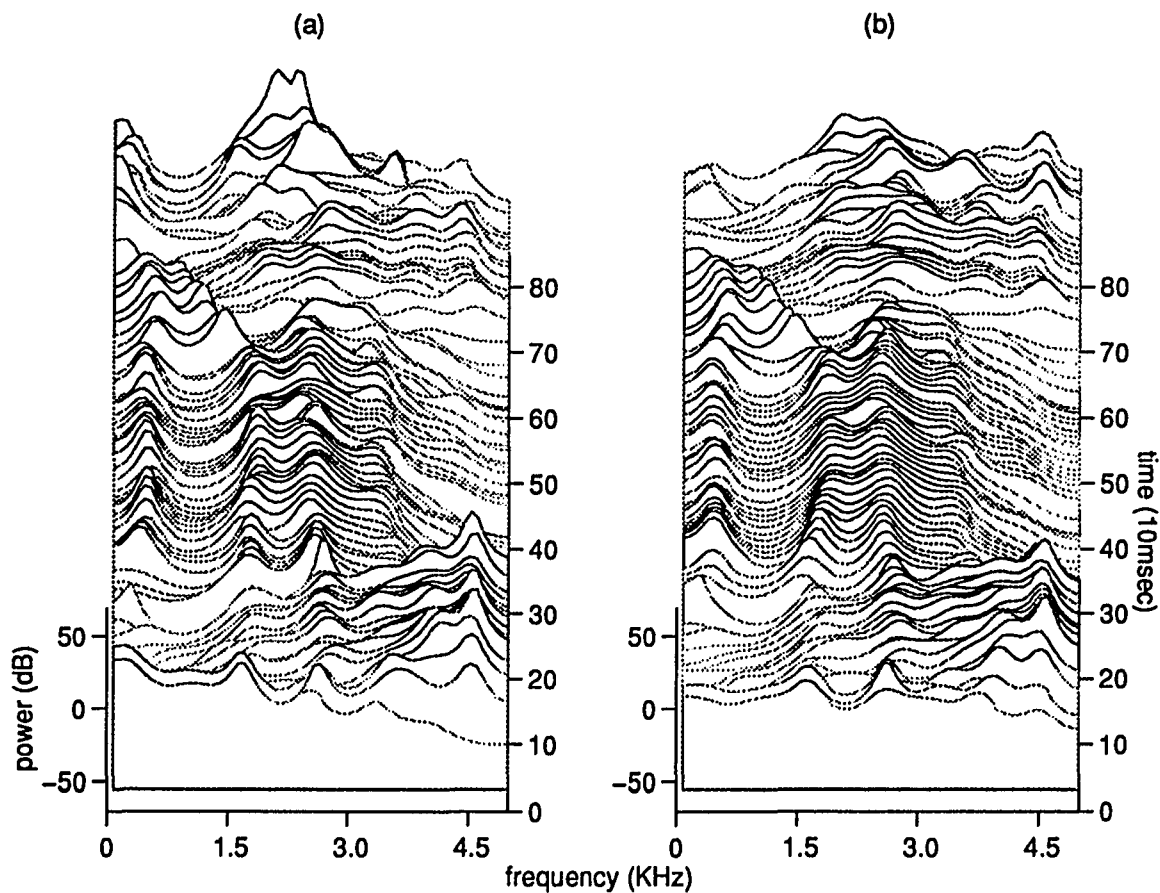


Figure 4.7: Illustration of LMR-based spectral conversion with and without overlapped training approach by three dimensional spectra of the word "sail": (a) by conventional LMR-based method, and (b) by LMR-based method with overlapped subset training approach

CHAPTER 5

APPLICATION TO ALARYNGEAL SPEECH ENHANCEMENT

The modified VQ- and LMR-based speech conversion methods were used for the enhancement of alaryngeal speech. A perceptual evaluation was completed to assess different methods of speech conversion and to determine if enhancement of alaryngeal speech was achieved using the modified speech conversion methods.

5.1 Subjects and Recordings

One laryngectomized male using tracheoesophageal speech and one laryngectomized female using tracheoesophageal speech provided the data. Both were proficient talkers who have used their methods of alaryngeal speech for a minimum of one year. Both were referred to this project by the speech-language pathologist responsible for their speech rehabilitation treatment, and were rated average to above-average in overall speech proficiency by their referring specialist. One male and one female normal talker provided the data for developing the conversion systems.

Recordings were made of subjects producing words and sentences (C.I.D. Auditory Test W-1, California Consonant Test Items, and Competing-Sentence Test) at a comfortable level of pitch and loudness. The recordings (SONY, TCD-D3) were

made in a quiet room with the recording microphone (ASTATIC, TM-80) placed about 5cm from the mouth of each talker. The recorded words were digitized into a computer at a sampling frequency of 10kHz (AT&T, DSP32-VME). The signal was passed through a low-pass filter (TTE, J73E) with a cut-off frequency of 4.5kHz prior to digitization. All subjects read the C.I.D. Auditory Test W-1 and California Consonant Test Items twice, and the Competing-Sentence Test once. The first list of the recorded words and sentences were used for system learning, and the second list of the recorded words were used for conversion and perceptual evaluation.

The recorded alaryngeal speech samples were analyzed before being used for system development and evaluation. The results of the analysis indicated that the vowel space of the male alaryngeal talker was noticeably different from normal speech, whereas the vowel space of the female talker was similar to normal speech. The spectra of the three "corner" vowels pronounced by the male alaryngeal and normal talkers are shown in Figure 5.1. The first and second formants of these vowels are plotted in Figure 5.2. In addition to the location of formants, the high frequency components are also stronger in the alaryngeal speech than in the normal speech. By contrast, no significant differences in formants were found between the female alaryngeal talker and the normal female talker.

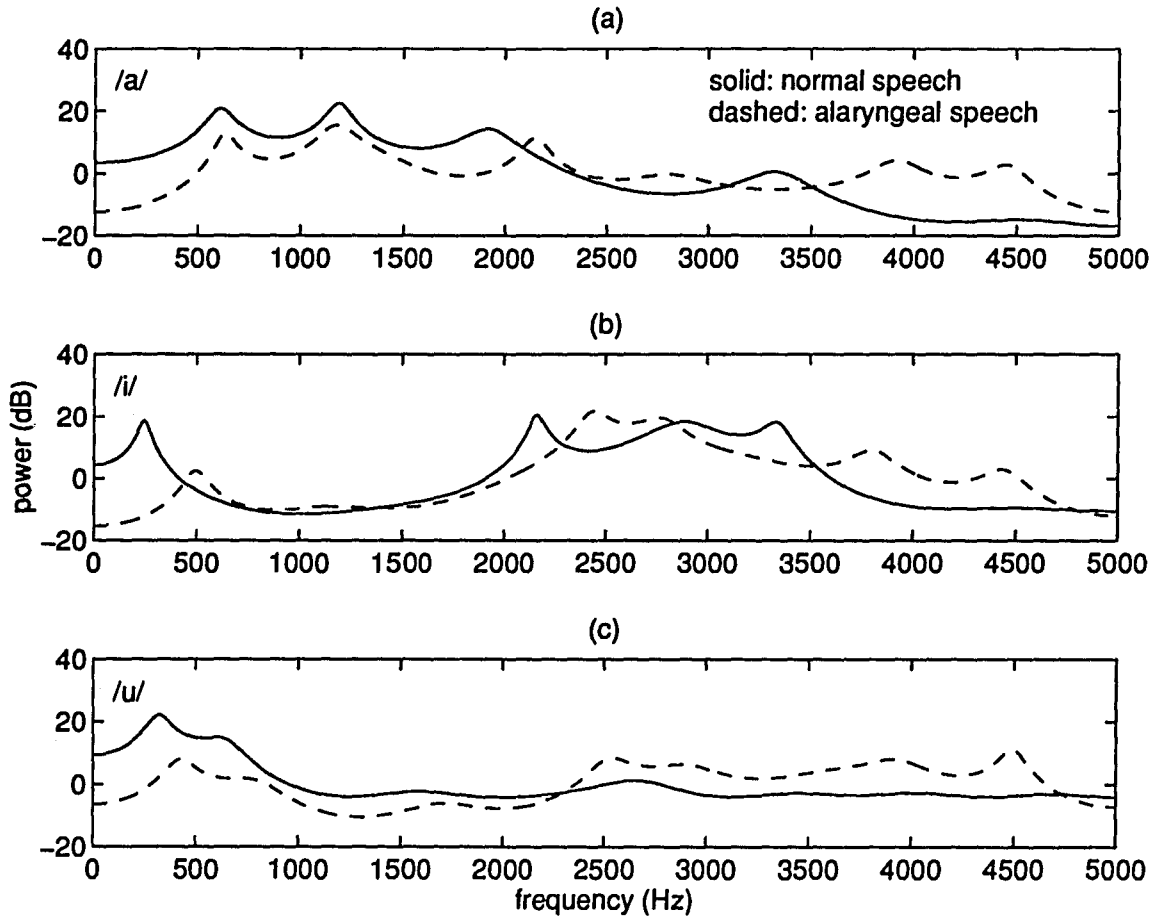


Figure 5.1: Formants of the male alaryngeal subject and the normal subject

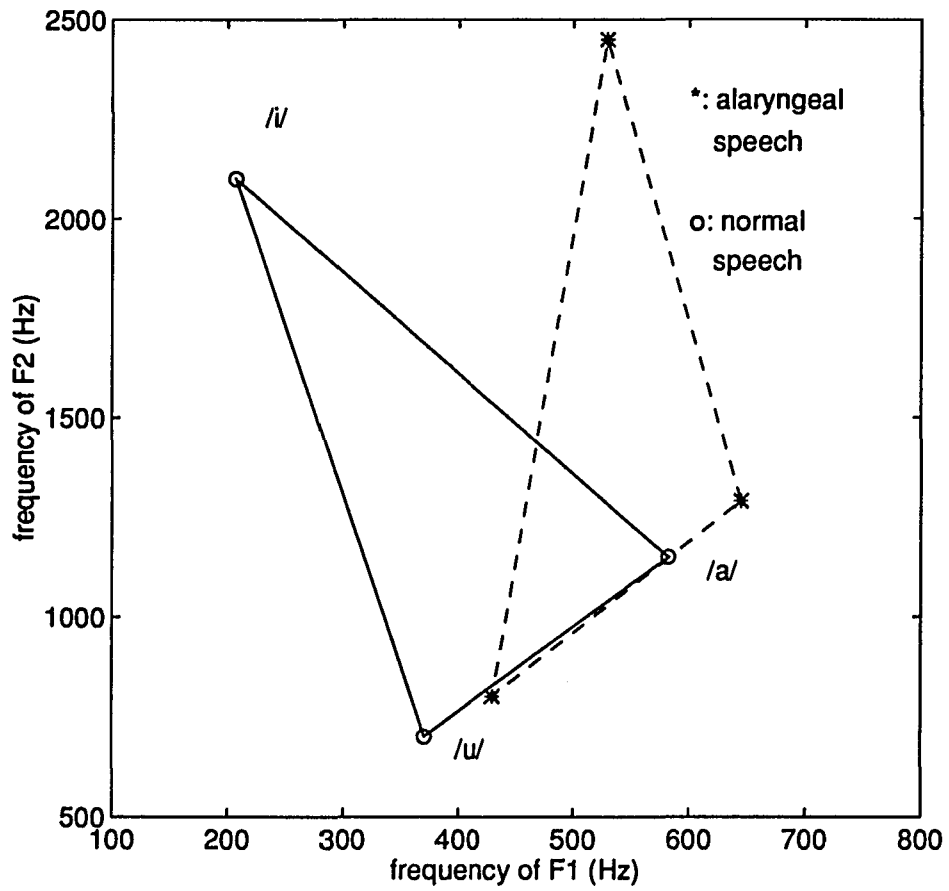


Figure 5.2: Vowel formant space of the alaryngeal subject and the normal subject

5.2 System Implementation

The speech conversion system has four major components: speech analysis, voice source replacement, spectral conversion, and speech synthesis. The implementation of each component is described as follows.

5.2.1 Speech Analysis

Speech signals were analyzed to obtain LPC coefficients. Only the voiced segment of each utterance was analyzed. A signal segment (or frame) was considered to be voiced when the fundamental period could be determined. Fourteen LPC coefficients were computed for each voiced frame using the auto-correlation method (see equation 2.18). Hamming window and pre-emphasis were used in the LPC analysis. Frame length was set to 40msec, and frame step-size was set to the current fundamental period. The LPC coefficients were transformed into cepstral coefficients with order 26 for spectral conversion and synthesis (see Figure 2.13).

5.2.2 Voicing Source Replacement

The cepstrum of the speech signal was also used to estimate the voiced segment period. A speech segment was considered to be voiced if its cepstral peak exceeded a pre-set threshold. The threshold of voicing determination was 0.1 for normal speech. The threshold of voicing determination was 0.05 for alaryngeal speech because of the weak periodicity in alaryngeal speech. The period was computed from the cepstral

peaks. The window length for computing the cepstrum was set to $51.2msec$ to include two or more periods for period determination.

Examples of period determination for an utterance are shown in Figure 5.3. Figure 5.3 (a) is the cepstra for a normal talker, Figure 5.3 (b) is the cepstra for an alaryngeal talker when the analyzing window was $25.6msec$. Figure 5.3 (c) is the cepstra for the same alaryngeal talker when the analyzing window was $51.2msec$. The periods computed from cepstral peaks were smoothed using a median filter.

The synthetic voicing excitation was generated based on the approximation of the LF-model (Qi and Bi, 1994). The temporal parameters of the LF-model, t_e , t_p , and t_c , were defined as a constant proportion of the period. Amplitude, E_e , was computed from the gain constant of the LPC filter (see Chapter 2 for details).

5.2.3 Spectral Conversion

5.2.3.1 VQ-Based Conversion System

The implementation of a VQ-based conversion system has two phases: the learning phase and the conversion-synthesis phase. In the learning phase, a mapping codebook that specifies the mapping function from the input spectral space to the target spectral space was generated. In the conversion-synthesis phase, speech signals were analyzed and, then, synthesized using the converted spectral transfer function.

In the learning phase, the same list of words and sentences produced by alaryngeal talkers and their normal target talkers were analyzed every $5msec$. The resulting

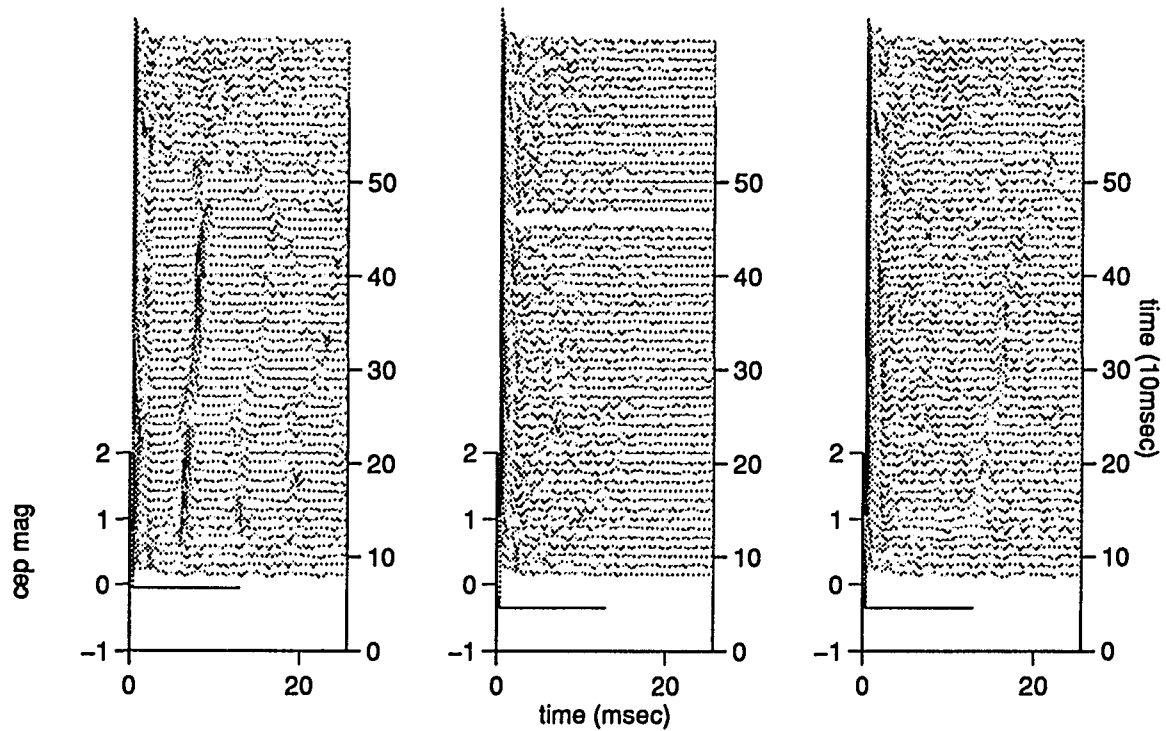


Figure 5.3: Period determination by cepstral coefficients: (a) normal speech, (b) alaryngeal speech with analysis window length of 25.6msec, and (c) alaryngeal speech with analysis window length of 51.2msec

input and target spectral-vectors were paired using procedure of DTW alignment (see Chapter 2 for details). Because the duration segments of alaryngeal speech often are longer than that of normal speech, the warping region was adjusted adaptively to accommodate the patterns to be matched. Assuming M and N are the durations of two spectral patterns and $M > N$, the slope of the top and bottom sides of the warping parallelogram was set to $\frac{N}{2M}$ instead of a fixed $\frac{1}{2}$ while the slope of the left and right sides was kept at 2 (the dashed line shown in Figure 2.10). If M were smaller than N , the slope of the left and right sides of the warping parallelogram was set to $\frac{2N}{M}$ instead of a fixed 2 while the slope of the top and bottom sides was kept at $\frac{1}{2}$. This adaptive modification of the warping region enabled the DTW algorithm to align most of the speech samples. The DTW total cost, $D(C)$ (see equation 2.51), was used as a parameter to identify speech samples which time-alignment was not possible. These samples were removed from system training.

Examples of DTW alignment are shown in Figures 5.4. Figures 5.4 (a) and (b) are the three-dimensional spectra for the word “sail” before DTW alignment. Figures 5.4 (c) and (d)) are the spectra for the same word after DTW alignment.

Given the input and target vectors and the pairing relations, the mapping codebook was obtained in three steps:

1. an input codebook, the codebook of input vectors, was obtained using vector quantization;

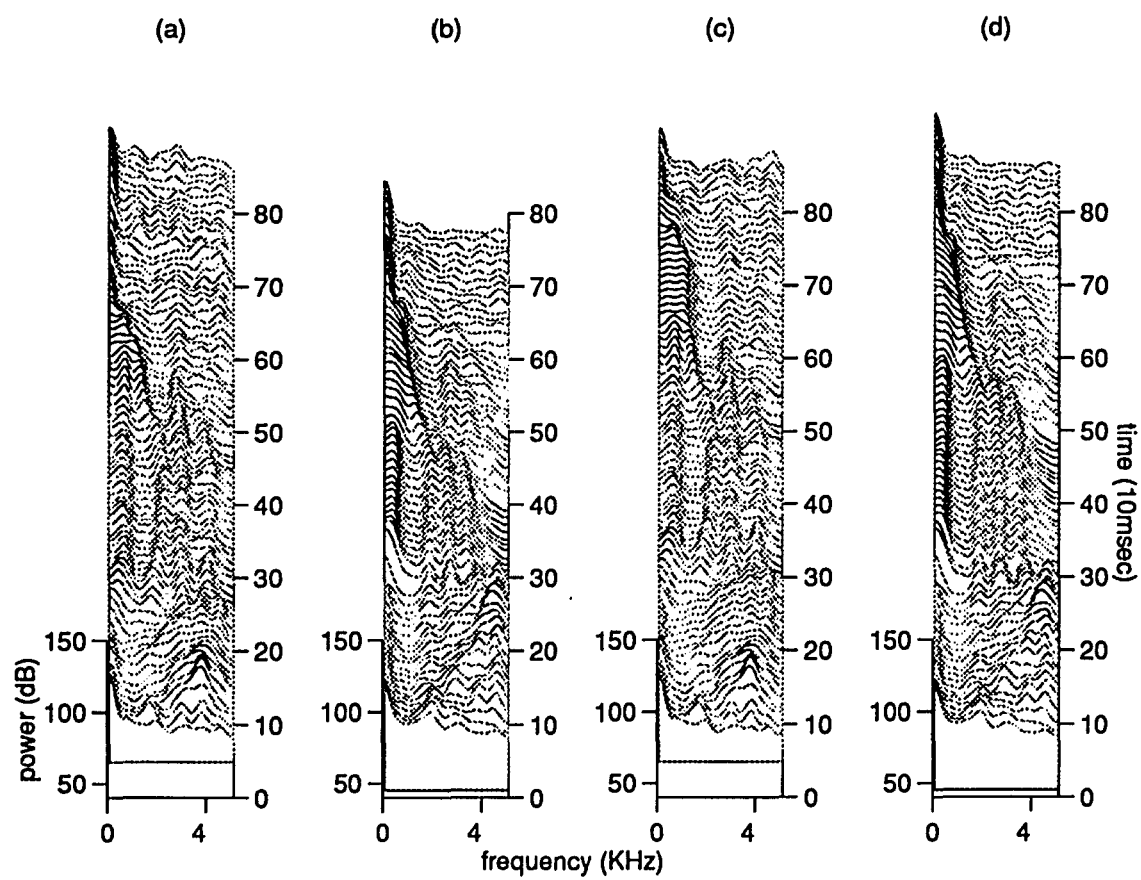


Figure 5.4: Illustration of DTW processing by three dimensional spectra: (a) and (b) original speech patterns, (c) and (d) after DTW matching

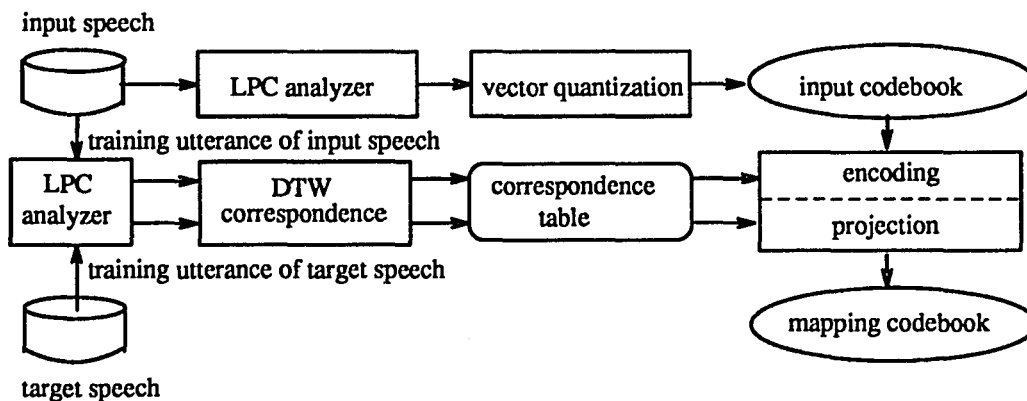


Figure 5.5: Block diagram of learning phase in the VQ-based spectral conversion approach

2. the projections (target vectors) from an input cluster were identified based on the pairing relations;
3. the average of these projections was designated to be the target codeword for the codeword of this input cluster.

This process is illustrated in Figure 5.5.

For comparison, the fuzzy vector quantization (FVQ) was also implemented to compute the mapping codebook. In this approach, each input vector had memberships in multiple input clusters. The membership was a function that is inversely proportional to the distance between the input vector and the cluster centers (see equation 3.3). The mapping codeword was computed again as a weighted average of all projections from a given input cluster. The weights were equal to the memberships. In both implementations, the size of the input codebook was set to 512. Consequently, the size of the target codebook was also 512.

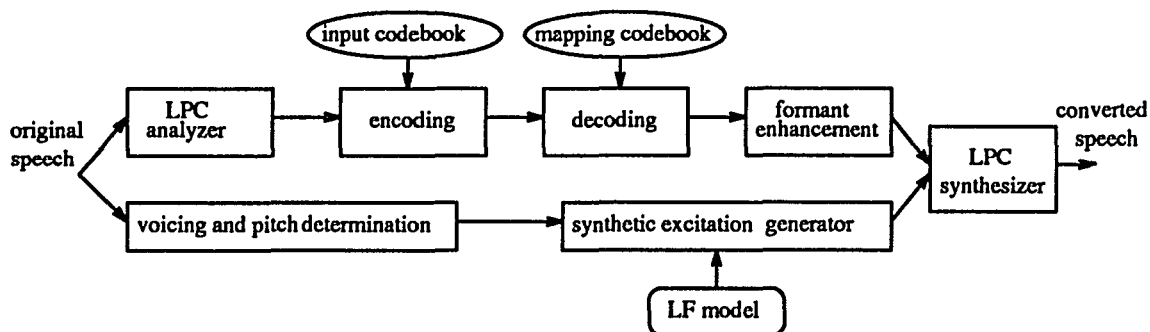


Figure 5.6: Block diagram of conversion-synthesis phase in the VQ-based spectral conversion approach

In the conversion-synthesis phase, an input frame of signal was analyzed and its cepstral coefficients were obtained. The input codeword for the cepstral coefficients were identified and conversion was made based on the mapping codebook. To enhance the formants, the converted cepstral coefficients were weighted by the sine window before being transformed into system impulse response (see equation 4.3, where h was set to 0.4). The impulse response was weighted again by the sequence, r^{-n} , $r = 0.98$ (see equation 4.5), to further enhance the formants. The impulse response was transformed to LPC coefficients. A period of speech signal was then synthesized through convolution between this impulse response and an excitation input. A block diagram of the conversion-synthesis process is illustrated in Figure 5.6.

5.2.3.2 LMR-Based Conversion System

The implementation of LMR-based conversion also involves a learning phase and a conversion-synthesis phase. In the learning phase, a mapping matrix that specifies

the mapping function from the input spectral space to the target spectral space was generated. In the conversion-synthesis phase, speech signals were analyzed and, then, synthesized using the converted spectral transfer function.

In the learning phase, the mapping matrix was also generated from pairs of input and target spectral vectors. These vectors were obtained using the same procedures as described in the previous section. Given the input and target vectors and the pairing relations, the mapping matrix was obtained as follows:

1. an input codebook, the codebook of input vectors, of a few clusters (64) was obtained using vector quantization;
2. the projections of each input cluster were identified based on the pairing relations;
3. a mapping matrix was computed using least-square approximations (see equation 3.10). The data sets for the least-square approximation were the vectors in the input cluster and their projections.

This process is illustrated in Figure 5.7.

In the conversion-synthesis phase of LMR-based system, an input spectrum is classified by the input codebook, and then is converted using the relative mapping matrix. A block diagram of the LMR-based system is shown in Figure 5.8.

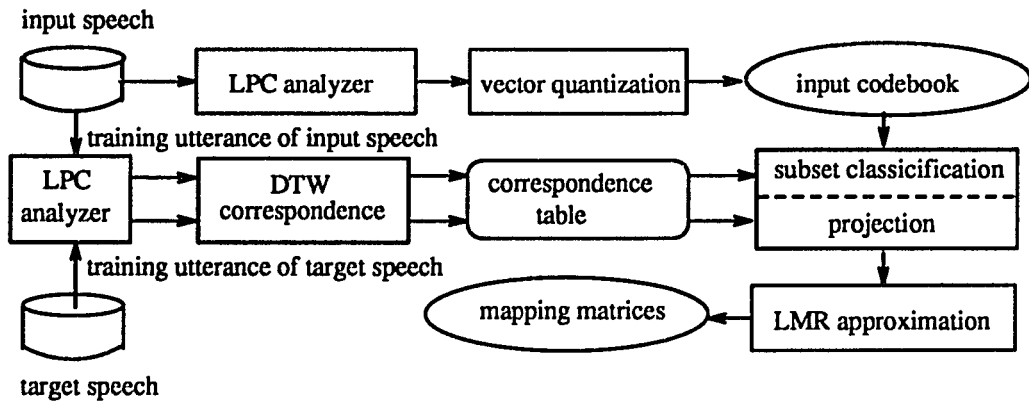


Figure 5.7: Block diagram of learning phase in the LMR-based spectral conversion approach

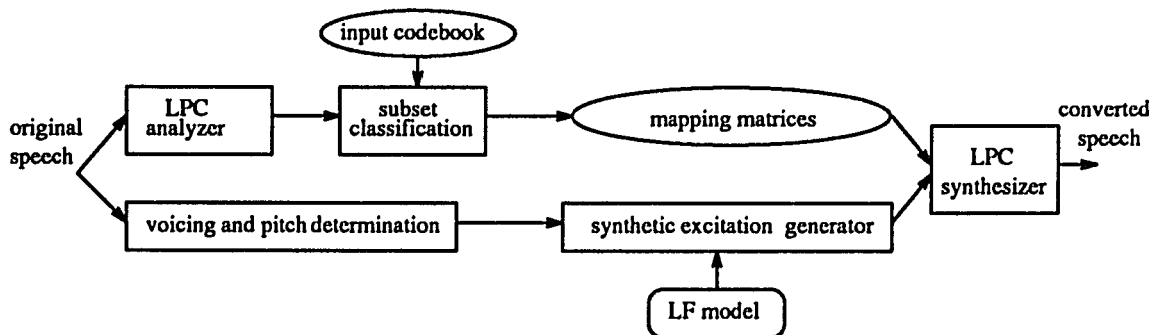


Figure 5.8: Block diagram of conversion-synthesis phase in the LMR-based spectral conversion approach

5.3 Perceptual Evaluation

A paired comparison approach was used to determine whether enhancement of alaryngeal speech was achieved using speech conversion systems. Six words (beach, drawbridge, inkwell, peep, sail, woodwork) produced by the alaryngeal talkers were selected for perceptual evaluation. These words were chosen because they provided a reasonably representative sampling of the vowel space.

Each word was synthesized under five conditions:

1. only the voicing source was replaced;
2. both the voicing source and the spectrum were replaced, and spectral conversion was made using the modified VQ-based conversion method;
3. both the voicing source and the spectrum were replaced, and spectral conversion was made using the modified LMR-based conversion method;
4. both the voicing source and the spectrum were replaced, and spectral conversion was made using the conventional VQ-based conversion method;
5. both the voicing source and the spectrum were replaced, and spectral conversion was made using the conventional LMR-based conversion method.

Each original word and its (1) to (3) synthetic counterparts were paired in all possible combinations. Conditions (2) and (4), and (3) and (5) were also paired,

respectively. All pairs were presented to the listeners. The order of the pairs in the presentation list was randomized.

Twelve students at the University of Arizona provided the preference judgments. Each listener was allowed to listen to each pair of words as many times as needed before determining which word in the pair "sounded more natural or was more pleasant to listen to." Each listener was also asked to listen to word pairs a second time. The order of the pairs in the list was re-randomized for the second presentation.

5.4 Evaluation Results

5.4.1 Listener Reliability

The reliability of listeners was evaluated by calculating the percentages of agreement in preference judgments made by each listener in response to the repeated presentation of all word pairs. The responses of listeners exhibiting 50% or greater test-retest agreement in preference judgments were used to evaluate enhancement. Ten listeners achieved this arbitrarily established criteria.

5.4.2 Summary of Evaluation Scores

The listeners' judgments of preference made in response to words synthesized by different enhancement systems, and original word produced by the male, alaryngeal

word pair characteristics	alaryngeal speech	only source replaced	modified VQ-based	VQ-based	LMR-based
only source replaced	68% 82/120				
modified VQ-based	82% 98/120	87% 104/120		64% 77/120	
modified LMR-based	80% 96/120	85% 102/120	50% 60/120		62% 74/120

Table 5.1: Number and percentage of responses preferring conditions of word in the first column (subject 1)

talker are summarized in Table 1. The data in Table 1 are the number and percentage of listeners preferring words synthesized under conditions described in the first column.

Based on a binomial distribution table (MacKinnon, 1964), these data reveal a significant ($p < 0.01$), clear overall preference by the listeners for the synthesized versions of words, demonstrating that enhancement of speech produced by this male laryngectomized talker was accomplished using speech analysis-synthesis methods with or without spectral conversion.

The data in Table 1 also revealed the impact of spectral conversion. Listeners preferred converted words over the words synthesized by replacing voicing source only. As expected, both the modified LMR- and VQ-based speech conversion approaches

word pair characteristics	alaryngeal speech	only source replaced
only source replaced	96% 159/165	
modified VQ-based	90% 148/165	43% 71/165

Table 5.2: Number and percentage of responses preferring conditions of word in the first column (subject 2)

achieved better performances than the conventional systems. The modified LMR-based method and the VQ-based method had comparable performance.

For the female alaryngeal talker, speech enhanced by the LPC analysis-synthesis method had the highest scores (see Table 5.2). Listeners almost unanimously preferred synthesized version of words over the originals. Listeners also preferred the speech samples synthesized by LPC system without spectral conversion.

These results indicated that speech conversion would be useful for alaryngeal talkers with articulatory deficits. The speech conversion would not be necessary when articulatory deficits are minimal.

CHAPTER 6

SUMMARY AND CONCLUSION

In this investigation, the original VQ- and LMR-based spectral conversion methods were modified. The modifications were aimed at reducing the spectral distortion in the VQ-based method and the spectral discontinuity in the LMR-based method. The modified systems were used for alaryngeal speech enhancement. Perceptual evaluations were completed to determine if enhancement could be accomplished using these modified speech conversion methods.

The spectral distortion (bandwidth increase) in the VQ-based speech conversion system is intrinsic to the algorithms of vector quantization and VQ mapping. Speech synthesized with large bandwidth sounds ambiguous and unclear. Because spectral averaging cannot be avoided in the VQ-based spectral conversion system, formant enhancement was included in the speech conversion process to compensate for the bandwidth increase. Formant enhancement was made using the chirp z-transform and the cepstral weighting method.

In the LMR-based approach, the spectral space was partitioned by a few large clusters and the spectrum within each cluster was mapped continuously; however, the transitions between clusters can be discontinuous resulting in audible clicks in the converted speech output. This discontinuity is largely due to the non-overlapped

subset is used to derive the LMR mapping matrix. To reduce the spectral discontinuity, a training method using overlapped subsets was developed. The advantage of using overlapped subsets during training is that the mapping matrix of each subset is constrained by samples of the subset and its neighboring subsets so that continuity between transitions can be maintained.

Using simulated data and natural speech, it was found that the LMR-based conversion system was more accurate than the VQ-based conversion system when the data set for system learning was large. The LMR-based system, however, was not stable when the data set for system learning was small. The VQ-based system appeared to be able to tolerate a limited training set. Thus, the LMR-system is recommended when the training data set is large. When the training data set is small, the VQ-based system is preferable.

Finally, speech conversion systems were implemented using the modified speech conversion methods. Results of perceptual evaluations indicated that listeners generally preferred the output of modified algorithms. The enhancement achieved by the modified LMR-based approach was comparable to that of the modified VQ-based approach. Results of perceptual evaluations also revealed that speech conversion techniques were more effective on alaryngeal speech with articulatory deficits.

The techniques developed in this investigation can be used in a broad range of applications. For example, they can be used in speech synthesis system to generate speech with different personalized voice characteristics. They may also be used in

secure speech communication system, speech coding system, and speech recognition system. They are potentially useful for the enhancement of alaryngeal and other types of disordered speech when articulatory deficits are the primary source of problem.

REFERENCES

- Abe, M. (1991). A segment-based approach to voice conversion. In *Proc. of IEEE International Conference on Acoust., Speech, Signal Processing*, volume ICASSP-91, pages 765-768, Toronto.
- Abe, M., Nakamura, S., Shikano, K., and Kuwabara, H. (1988). Voice conversion through vector quantization. In *Proc. of IEEE International Conference on Acoust., Speech, Signal Processing*, volume ICASSP-88, pages 655-658, New York.
- Abe, M., Shikano, K., and Kuwabara, H. (1990). Cross-language voice conversion. In *Proc. of IEEE International Conference on Acoust., Speech, Signal Processing*, volume ICASSP-90, pages 345-348.
- Ananthapadmanabha, T. (1984). Acoustic analysis of voice source dynamics. *STL-QPSR*, 2-3:1-24.
- Ananthapadmanabha, T. and Fant, G. (1982). Calculation of true glottal flow and its components. *STL-QPSR*, 1:1-30.
- Atal, B. and Schroeder, M. (1967). Predictive coding of the speech signals. *Proc. 1967 Conf. Commun. and Process.*, pages 360-361.

- Bennett, S. and Weinberg, B. (1973). Acceptability ratings of normal, esophageal, and artificial larynx speech. *Journal of Speech and Hearing Research*, 38:608–615.
- Childers, D. and Lee, C. (1991). Vocal quality factors: Analysis, synthesis, and perception. *J. Acoust. Soc. Amer.*, 90:2394–2410.
- Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton, S-Gravenhage.
- Fant, G. (1981). The source filter concept in voice production. *STL-QPSR*, 1:21–37.
- Fant, G., Liljencrants, J., and Lin, Q. G. (1985). A four-parameter model of glottal flow. *STL-QPSR*, 4:1–12.
- Flanagan, J. L. (1972). *Speech Analysis Synthesis and Perception*. Springer-Verlag, Berlin Heidelberg New York, 2nd edition.
- Gray, A. and Markel, J. (1976). Distance measures for speech processing. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-24:381–391.
- Ishizaka, K. and Flanagan, J. (1972). Synthesis of voice sounds from a two-mass model of the voice cords. *Bell System Technical Journal*, 51:1233–1268.
- Itakura, F. (1965). Minimum prediction residual principle applied to speech recognition. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-23:67–72.
- Klatt, D. and Klatt, L. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Amer.*, 87:820–857.

- Kohonen, T. (1977). *Associative Memory*. Springer-Verlag, Berlin Heidelberg New York, 1st edition.
- Kohonen, T. (1989). *Self-Organization and Associative Memory*. Springer-Verlag, Berlin Heidelberg New York London Paris Tokyo Hong Kong, 3rd edition.
- Linde, Y., Buzo, A., and Gray, R. (1980). An algorithm for vector quantizer design. *IEEE Trans. Commun.*, COM-28:84–95.
- MacKinnon, W. (1964). Table for both the sign test and distribution free confidence intervals of the median for sample sizes to 1000. *Journal of American Statistical Association*, 59:935–956.
- Markel, J. and Gray, A. (1976). *Linear Prediction of Speech*. Springer-Verlag, Berlin Heidelberg New York.
- McCandless, S. (1974). An algorithm for automatic formant extraction using linear predictive spectra. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-22:135–141.
- Nakamura, S. and Shikano, K. (1989). Speaker adaptation applied to hmm and neural networks. In *Proc. of IEEE International Conference on Acoust., Speech, Signal Processing*, volume ICASSP-89,S3.3, pages 89–92.
- Nemhauser, G. (1966). *Introduction to Dynamic Programming*. Wiley, New York.
- Noll, A. (1967). Cepstrum pitch determination. *J. Acoust. Soc. Amer.*, 41:293–309.

- Nord, L. and Hammarberg, B. (1989). Analysis of laryngectomee speech – a progress report. In *European Conference on Speech Communication and Technology*, volume 2, pages 493–496.
- Oppenheim, A. and Schaffer, R. (1989). *Discrete -Time Signal Processing*. Prentice-Hall, Englewood Cliffs, New Jersey.
- Parsons, T. (1987). *Voice and Speech Processing*. McGraw-Hill, New York.
- Qi, Y. (1990). Replacing tracheoesophageal voicing sources using LPC synthesis. *J. Acoust. Soc. Amer.*, 88:1228–1235.
- Qi, Y. and Bi, N. (1994). A simplified approximation of the four-parameter LF model of voice source. *J. Acoust. Soc. Amer.*, 96:1182–1185.
- Qi, Y. and Weinberg, B. (1991). Spectral slope of vowels produced by tracheoesophageal speakers. *Journal of Speech and Hearing Research*, 34:243–247.
- Qi, Y. and Weinberg, B. (1995). Characteristics of voicing source waveforms produced by esophageal and tracheoesophageal speakers. *Journal of Speech and Hearing Research*, 38:In Press.
- Qi, Y., Weinberg, B., and Bi, N. (1995). Enhancement of female esophageal and tracheoesophageal speech. *J. Acoust. Soc. Amer.*, In Review.
- Rabiner, L. and Juang, B. (1993). *Fundamentals of Speech Recognition*. Prentice-Hall, New Jersey.

- Rabiner, L., Levinson, S., and Sondhi, M. (1983). On the application of vector quantization and hidden markov models to speaker-independent, isolated word recognition. *Bell System Tech. J.*, 62:1075–1105.
- Rabiner, L., Rosenberg, A., and Levison, S. (1978). Considerations in dynamic time warping algorithms for discrete word recognition. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP–26:575–582.
- Rabiner, L. and Schafer, R. (1978). *Digital Processing of Speech Signals*. Prentice-Hall, New Jersey.
- Rabiner, L., Schafer, R., and Rader, C. (1969). The chirp z-transform algorithm. *IEEE Trans. on Audio and Electroacoustics*, AU–17:86–92.
- Robbins, J., Fisher, H., Blom, E., and Singer, M. (1984). A comparative acoustic study of normal, esophageal and tracheoesophageal speech production. *Journal of Speech and Hearing Research*, 49:202–210.
- Satio, S. and Itakura, F. (1966). *The Theoretical Consideration of Statistically Optimum Methods for Speech Spectral Density*. Report No. 3107. Electrical Communication Laboratory, N.T.T., Tokyo.
- Savic, M. and Nam, I. (1991). Voice personality transformation. *Digital Signal Processing*, 1:107–110.

- Shikano, K., Lee, K., and Reddy, R. (1986). Speaker adaptation through vector quantization. In *Proc. of IEEE International Conference on Acoust., Speech, Signal Processing*, volume ICASSP-86, pages 2643–2646, Tokyo.
- Shikano, K., Nakamura, S., and Abe, M. (1991). Speaker adaptation and voice conversion by codebook mapping. In *Proc. of IEEE International Symposium on Circuits and System*, volume 1, pages 594–597.
- Sisty, N. and Weinberg, B. (1972). Vowel formant frequency characteristics of esophageal speech. *Journal of Speech and Hearing Research*, 15:439–448.
- Smith, B., Weinberg, B., Feth, L., and Horii, Y. (1978). Vocal roughness and jitter characteristics of vowels produced by esophageal speakers. *Journal of Speech and Hearing Research*, 21:240–249.
- Stevens, K. (1989). On the quantal nature of speech. *J. of Phonetics*, 17:3–46.
- Stevens, K. and House, A. (1955). Development of a quantitative description of vowel articulation. *J. Acoust. Soc. Amer.*, 27:484–493.
- Trudeau, M. and Qi, Y. (1990). Acoustical characteristics of female tracheoesophageal speech. *Journal of Speech and Hearing Disorders*, 55:244–250.
- Tseng, H., Sabin, M., and Lee, E. (1987). Fuzzy vector quantization applied to hidden markov modeling. In *Proc. of IEEE International Conference on Acoust., Speech, Signal Processing*, volume ICASSP-87,15.5, pages 641–644.

- Valbret, H., Moulines, E., and Tubach, J. (1992). Voice transformation using psola technique. *Speech Communication*, 11:175–187.
- Weinberg, B. (1982). Speech after laryngectomy: An overview of acoustic and temporal characteristics of esophageal speech. In Sekey, A. and Hanson, R., editors, *Electroacoustic Analysis and Enhancement of Alaryngeal Speech*, pages 5–48. Thomas Co., Springfield.
- Weinberg, B. (1986). Acoustical properties of esophageal and tracheoesophageal speech. In Keith, R. and Darley, F., editors, *Laryngectomy Rehabilitation*. College Hill Press, San Diego.
- Weinberg, B. and Bennett, S. (1972). A comparison of fundamental frequency characteristics measured on a wave-by-wave and averaging basis. *Journal of Speech and Hearing Research*, 15:351–355.
- Weinberg, B., Horii, Y., and Smith, B. (1980). Long time spectral and intensity characteristics of esophageal speech. *J. Acoust. Soc. Amer.*, 67:1781–1784.
- White, G. and Neely, R. (1976). Speech recognition experiments with prediction bandpass filtering, and dynamic programming. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-24:183–188.