Transformation of Vocal Tract Characteristics for Voice Conversion using Artificial Neural Networks

A THESIS
submitted for the award of the degree
of
MASTER OF SCIENCE
in
COMPUTER SCIENCE AND ENGINEERING

by
NARENDRANATH M.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECI-INOLOGY
MADRAS 600 036
November 1995

CERTIFICATE

This is to certify that the thesis entitled

"Transformation of Vocal Tract Characteristics for Voice Conversion using Artificial Neural Networks"

submitted by Narendranath. M to the Indian Institute of Technology, Madras for the award of the degree of Master of Science in Computer Science and Engineering is a bonafide record of research work carried out by him under my guidance and supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

(B. Yegnanarayana)

Madras 600 036

Date:

Acknowledgements

I thank my research supervisors Prof. B. Yegnanarayana and Dr. Hema A. Murthy for their cooperation and able guidance extended to me throughout this work. It was Prof. Yegnanarayana who introduced me to the field of speech signal processing and neural networks. His inspiring and enlightening lectures on signal processing and neural networks helped me in acquiring the fundamentals of these areas. I found in him an ideal teacher. I gratefully acknowledge the time that he has spent in going through the earlier drafts of this thesis. I thank him for his valuable comments and suggestions on my thesis work. Many of the research ideas presented in.this thesis have come out from such suggestions.

I also thank Dr. Hema A. Murthy for her constant help and motivation. I had a number of technical discussions with Dr. Hema from which I have benefitted a lot. I have extensively used her programs during the course of this work. I sincerely thank her for the time and effort spent by her in discussing with me various research issues.

Most of the work was supported by Department of Science and Technology, Govt. of India through the sponsored project "Voice conversion", at I.I.T. Madras. I am grateful for their support.

I extend my sincere thanks to V.R. Ramachandran who helped me in getting myself familiar with the computer systems and other equipment in the lab. My colleague S. Rajendran helped me in many ways during my stay at I.I.T Madras. He was a rich and generous source of linguistic knowledge. We had numerous technical discussions which helped in improving my understanding of the speech code. I should also thank him for reading and correcting drafts of the thesis. I enjoyed working with him and I thank him for all the help extended to me.

The stimulating research environment in the Speech and Vision Lab, has helped me enormously in building research skills. I thank all the members of the Speech and Vision Lab who helped me in one way or the other in completing this work. My friend P. S. Murthy helped me a lot in bringing the thesis upto this stage. He helped me by proof reading the thesis a number of times. I am thankful to him for giving me an excellent company through out my stay at I.I.T. My colleague P.P Raghu went through portions of the thesis and his comments have significantly improved the readability of the thesis. I also thank Krishnan, my lab mate, for his help during the final stages of thesis preparation.

I thank my cousin brother Athmanand and his wife Anitha for providing me with accommodation during my early days at I.I.T. I am thankful to them for all the care taken in making my stay comfortable. I also thank my friend Ramesh N. with whom I stayed during the finishing stages of the thesis work.

I thank my parents, M. Sushila and P. Achuthan and my brothers, Muralidharan and Thulasidas for giving valuable advice at critical points of time. I also thank them for providing me with an excellent atmosphere at home.

Contents

Abstract

Speech signal contains two kinds of information. They are: (i) The message the speaker wants to convey to the listener and (ii) the characteristics of the speaker. In this thesis we focus on the analysis and manipulation of speaker characteristics embedded in the speech signal for ***voice conversion.*** Voice conversion involves transformation of the speaker characteristics in the speech uttered by a speaker (source speaker), so as to generate speech having the voice characteristics of the desired speaker (target speaker). Voice characteristics lie at the linguistic, suprasegmental and segmental levels. The speaker characteristics at the linguistic and suprasegmental levels are learned features. Hence they are difficult to derive from data and model. Speaker characteristics at the segmental level can be attributed to the speech production mechanism and they are reflected in the source and system characteristics of the physical system. The interspeaker variations of the vocal tract system can be modeled as transformation operations. Speech synthesized from the transformed parameters reflect the voice characteristics of the target speaker. The present study focuses on the transformation of the vocal tract system characteristics between two speakers and incorporation of the transformed characteristics in a voice conversion system. A major issue in this transformation task is to arrive at a suitable representation of the vocal tract system. For this we have selected formants as they provide a good representation of the vocal tract shape and at the same time can be easily extracted from the speech data.

We first explore the possibility of using linear transformations for transforming formants corresponding to steady vocal tract shapes (such as vowels) between speakers. While testing we have observed that if we use a single linear transform the error in the transformed formants is high. We noted that this error in transforming formants can be significantly reduced by using piecewise linear transformations. But piecewise linear transformations have the disadvantage of introducing discontinuities while transforming transitions in the formants. This is because, even for steady vocal tract shapes, the scaling of formants between speakers is highly nonlinear. We have

explorecl the possibility of using a inultilayer feedforward neural network to capture these nonlinear transformations of the formants. Using proper training data it is possible to design a network to transform not only the steady formants but also the formant transitions in dynamic sounds. Issues involved in implementing these transformations in a voice conversion system are addressed. Finally, we present the performance of the system for converting speech from one voice to another.

The major contributions of the thesis are: (i) Interspeaker variations in the formant locations are analyzed to show that the formant transformation between two speakers is highly nonlinear. (ii) A neural network-based formant transformation scheme is developed which works well even for formant transitions occurring in continuous speech. (iii) A method for measuring the generalization capability of the resulting network is proposed. (iv) A method for modifying the linear predictive coefficients(LPCs) is proposed to incorporate the transformation of formants in a voice conversion system.

Chapter 1

Introduction

1.1 Objective of the study

The purpose of speech is communication (communicative intent) [1, 2]. We use speech for communicating a variety of messages. Speech signal also carries with it information other than the message which a speaker intends to convey to a listener. This information includes the identity of the speaker, his emotional state, his physical state etc. Human beings are able to recognize a familiar speaker effortlessly from his speech. The focus of our work is to extract the speaker specific information contained in the speech signal for ***voice conversion.*** Voice conversion involves transformation of the speaker characteristics in the speech uttered by a source speaker, so as to generate speech in the voice of the desired target speaker. For developing a voice conversion system one has to identify the speaker dependent features ancl represent them in a suitable form. This representation is used to transform the speaker dependent features extracted form the speech of the source speaker into the features of the target speaker and speech is then synthesized using the transformed features.

1.2 Background

Speech signal contains mainly two kinds of information. They are: (i) The message that the speaker intends to convey to the listener and (ii) the identity of the speaker. Extracting the message part from the speech signal is the focus of research in the area of speech recognition and speech understanding [3, 4]. The area of speaker recognition and verification deals with techniques to extract the speaker dependent information from the speech signal [5, 6, 7].

In the development of a voice conversion system, speaker dependent. knowledge is acquired in the analysis or learning phase. In the transformation phase, the acquired (target)speaker dependent knowledge is used to modify the speaker dependent parameters extracted from the speech of the source speaker. Finally, speech with the voice characteristics of the target speaker is synthesised using the transformed parameters. This is relevant in many situations. For example, in a text-to-speech system, it may be required to generate speech with some desired voice characteristics. Analysis of speaker dependent characteristics is also useful for developing speaker recognition and speaker verification systems in security and forensic applications. Understanding speaker dependent characteristics is useful in speaker normalization for speaker independent speech recognition systems [S].

Voice conversion could be speaker-dependent or speaker-independent. In both the cases identity of the target speaker is fixed. In a speaker-dependent voice conversion scenario, the source speaker is also fixed. The task is to transform the voice characteristics in the speech of the source speaker to that of the target speaker. In speaker-independent voice conversion the task is to transform the characteristics of any speaker, so that the transformed speech sounds like that of the target speaker. In a speaker-independent voice conversion scheme the number of source speakers are unlimited but identity of the target speaker is fixed. In this work we address only speaker dependent voice conversion.

The major issues involved in the development of a voice conversion system are:

(i) The characteristics of the desired voice have to be identified and specified. This involves acoustic-phonetic analysis of speech data for each speaker. (ii) The acquired speaker dependent knowledge must be represented in a form suitable for transformation of speech from the source voice to the target voice. This may be represented as transformations which can transform the speaker dependent parameters extracted from the speech of the source speaker to match with that of the target speaker. The voice characteristics of the target speaker can also be represented as a set of rules, to incorporate the desired speaker characteristics into the source speech. (iii) Finally voice conversion is achieved by incorporating features of the target speaker into the parameters extracted from the source speech, and then synthesizing speech.

For developing a voice conversion system we must identify the factors in the speech signal which are responsible for giving individuality to the speech of a speaker. Speaker characteristics exist at various levels. Figure 1.1 shows knowledge sources used at various levels for producing and perceiving voice characteristics.

At the highest level, namely the linguistic level, we use factors like, language, dialect, syntactic structures and semantic context for the identification of a speaker form his speech. The characteristics of a speaker at this level are difficult to analyze and model, although these characteristics are mainly used by humans for recognizing speakers from spontaneous speech.

There are factors at the acoustic level which can be extracted from the speech waveform. The acoustic level cliaracterization can be divided further into segmental and suprasegmental levels. At the suprasegmental level the prosodic features such as intonation, duration and stress carry significant speaker-specific information. After the linguistic factors the prosodic factors are the most important speaker-specific characteristics which human beings use in recognizing speakers. At the segmental level the source and system characteristics of the speech production mechanism reflect the speaker characteristics. The system characteristics refer to the shape and size (mainly the effective length) of the vocal tract. Source characteristics refer to the physiology of the vocal folds. The segmental speaker characteristics have a

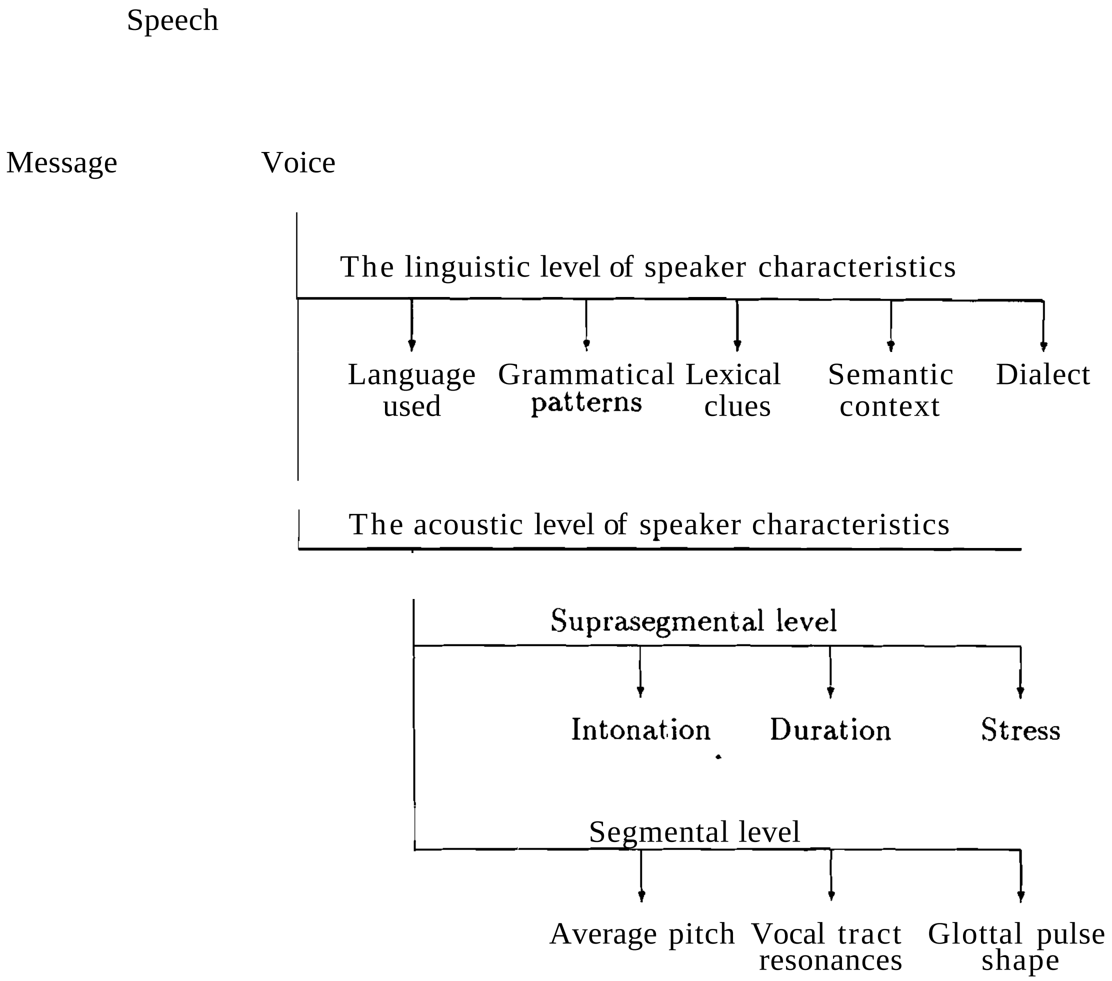


Figure 1.1: Line diagram showing the various speaker dependent knowledge sources.

dynamic and a static part. The dynamic part of the speech production contributes to speaker characteristics in the speech signal. This includes both the vocal tract system dynamics and the glottal source dynamics. These dynamic features are dictated by sound units and thus are determined by the text to a large extent. Static speaker characteristics refer to the average length of the vocal tract system, average pitch, characteristics of the nasal tract etc. are useful mainly for transforming steady sounds across speakers.

1.3 Scope of the thesis

Analysis and modeling of speaker characteristics at the linguistic and suprasegmental levels are difficult tasks. The speaker characteristics at the linguistic level gains importance when we deal with spontaneous speech. In this work we are concentrating only on read speech, where the speaker is asked to read aloud given sentences. Thus we eliminate speaker characteristics at the linguistic level from the speech data.

The interspeaker variations at the prosodic level can be attributed to several complex mental phenòmena (learning). These variations have no relation to any physical system. The prosodic characteristics of the speaker are derived by analysing large amount of speech data. This knowledge acquisition process involves significant manual effort.

The segmental characteristics are directly related to the physical system, namely the vocal tract system. Therefore at the segmental level features of the source speaker can be transformed into features corresponding to the target speaker. In this work we will be modeling the interspeaker variations at the segmental level across speakers as transformation functions. The emphasis is on transforming the characteristics of the vocal tract system. Only some gross features of source characteristics are considered in this thesis.

The problem of voice conversion using information at segmental level can be understood from the nature of the speech production mechanism and from the manifestation of the production characteristics in the speech signal. We now briefly present the fundamentals of speech production mechanism. The organs involved in the production of speech are shown in Figure 1.2. If the vocal tract system is excited by quasi-periodic vibrations of the vocal folds, then the resulting speech is called voiced (eg. vowels /a/ and /i/). The periodicity of the vocal cord vibration is called the pitch of the voice source. If the excitation of the vocal tract system is due to turbulence of air (frication) at a narrow constriction, the resulting speech

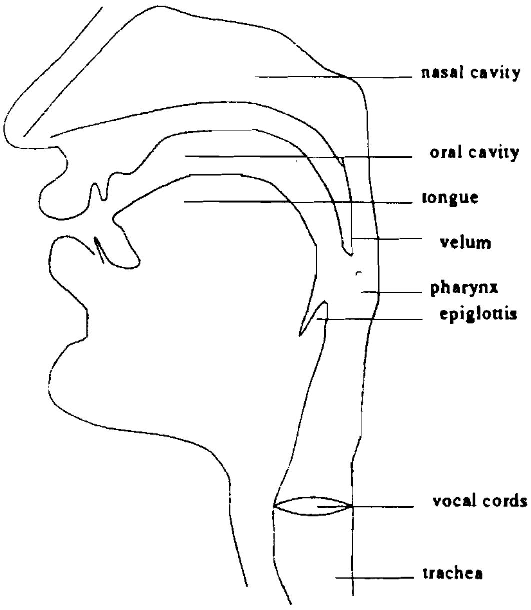


Figure 1.2: Organs of speech production

is said to be unvoiced (eg. /s/ and /z/). On the other hand if the vocal tract is closed at some point and the built up pressure is released suddenly the resulting speech is called a stop sound (eg. /p/ and /t/). Figure 1.3 shows a typical speech waveform where the three types of sounds are illustrated. During normal speech production the time varying source produces varying pitch frequency (F_0). This is called intonation contour or pitch contour as illustrated in Figure 1.4(a) and (b). The time varying vocal tract system characteristics are reflected as time varying resonances (formants) of the vocal tract system. These formant changes can be seen in a spectrographic display of speech signal as shown in Figure 1.4(e) and (f). Spectrogram is a display of the distribution of spectral information with respect to time. In a spectrogram time and frequency are represented in the x and y axes respectively while the amplitude is noted by the darkness of the picture. Formants appear as dark horizontal bands. Figure 1.4 shows the pitch contour and formant contour for two speakers uttering the vowel sequence /āī/. The problem of voice

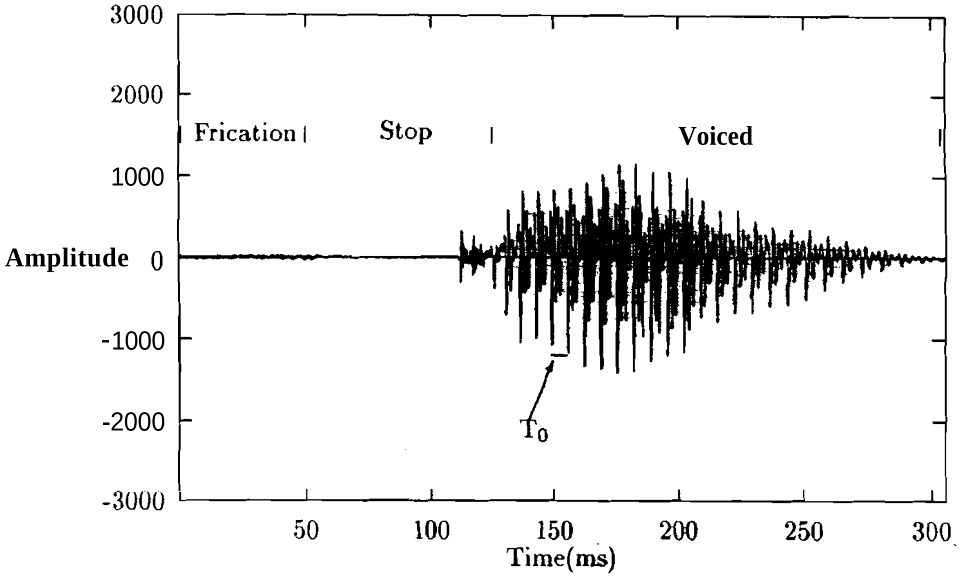


Figure 1.3: Waveform of the utterance 'sky'.

conversion using parameters at the segmental level is illustrated using the Figure 1.4. Figure 1.4(c) and (d) show the acoustic waveform of the speech sound /āī/ uttered by a male and a female speaker respectively. The F_0 contour extracted from these utterances are shown in Figure 1.4 (a) and (b). We can observe that the average F_0 of the male speaker is significantly lower than the F_0 of the female speaker. Thus, in order to perform a voice transformation across two voices, the average F_0 must be appropriately modified. Apart from the interspeaker variations in the source characteristics, the vocal tract system also contributes to speaker variability. The interspeaker variations in the vocal tract system are manifested as variations in the formant frequency (vocal tract resonances). Figure 1.4 (e) and (f) illustrates this with the help of spectrograms. The dynamics of the vocal tract system is manifested in the spectrogram in the form of smooth formant transitions. By comparing Figure 1.4(e) and (f), we observe that the formant frequencies is significantly.higher for female speech in comparison with male speech. Hence we note that for realising voice conversion, the formants extracted from the speech of the source speaker have to be appropriately transformed. Our aim is to capture a transformation operation

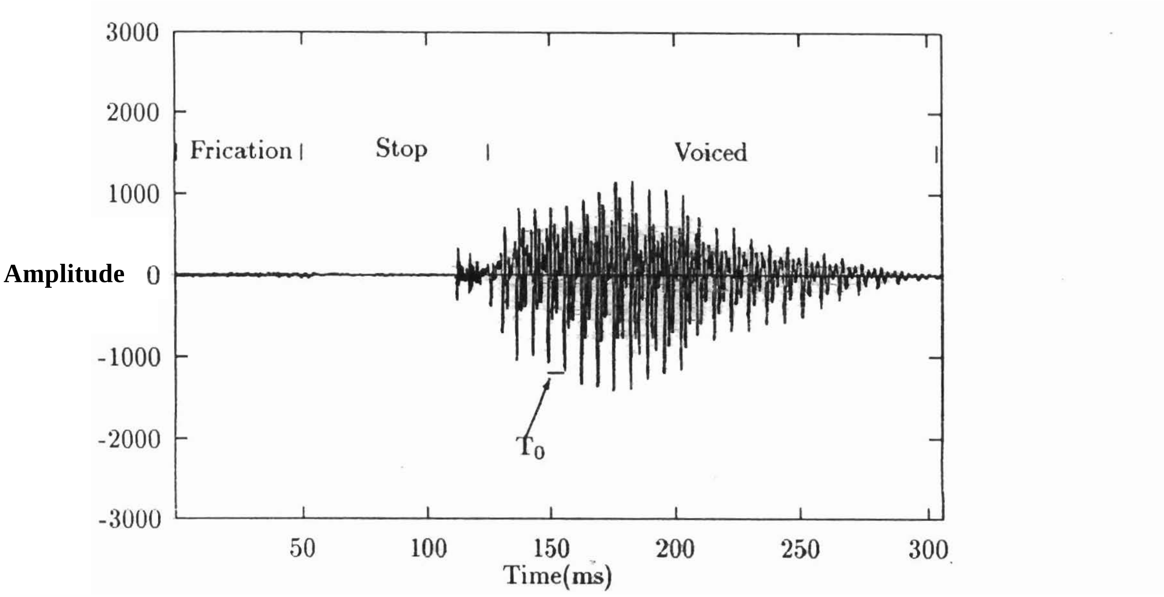


Figure 1.3: Waveform of the utterance 'sky'.

conversion using parameters at the segmental level is illustrated using the Figure 1.4. Figure 1.4(c) and (d) show the acoustic waveform of the speech sound /$\bar{a}\bar{\imath}$/ uttered by a male and a female speaker respectively. The F_0 contour extracted from these utterances are shown in Figure 1.4 (a) and (b). We can observe that the average F_0 of the male speaker is significantly lower than the F_0 of the female speaker. Thus, in order to perform a voice transformation across two voices, the average F_0 must be appropriately modified. Apart from the interspeaker variations in the source characteristics, the vocal tract system also contributes to speaker variability. The interspeaker variations in the vocal tract system are manifested as variations in the formant frequency (vocal tract resonances). Figure 1.4 (e) and (f) illustrates this with the help of spectrograms. The dynamics of the vocal tract system is manifested in the spectrogram in the form of smooth formant transitions. By comparing Figure 1.4(e) and (f), we observe that the formant frequencies is significantly higher for female speech in comparison with male speech. Hence we note that for realising voice conversion, the formants extracted from the speech of the source speaker have to be appropriately transformed. Our aim is to capture a transformation operation

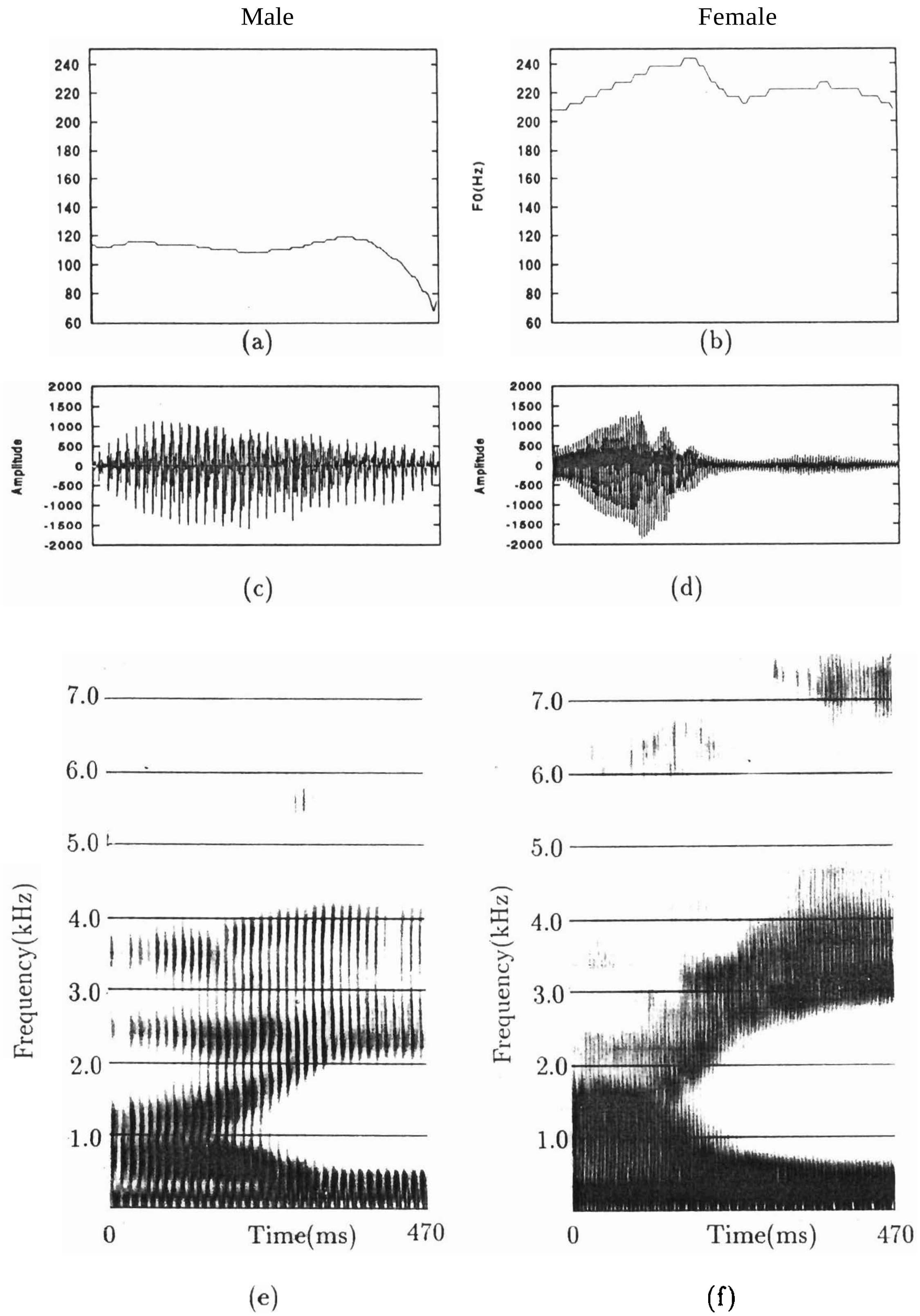


Figure 1.4: Illustration of the interspeaker variation at the segmental level. (c) and (d) shows the speech waveform for the sound /āī/. (a) and (b) show the pitch contours extracted from these utterances. (e) and (f) show the corresponding spectrograms.

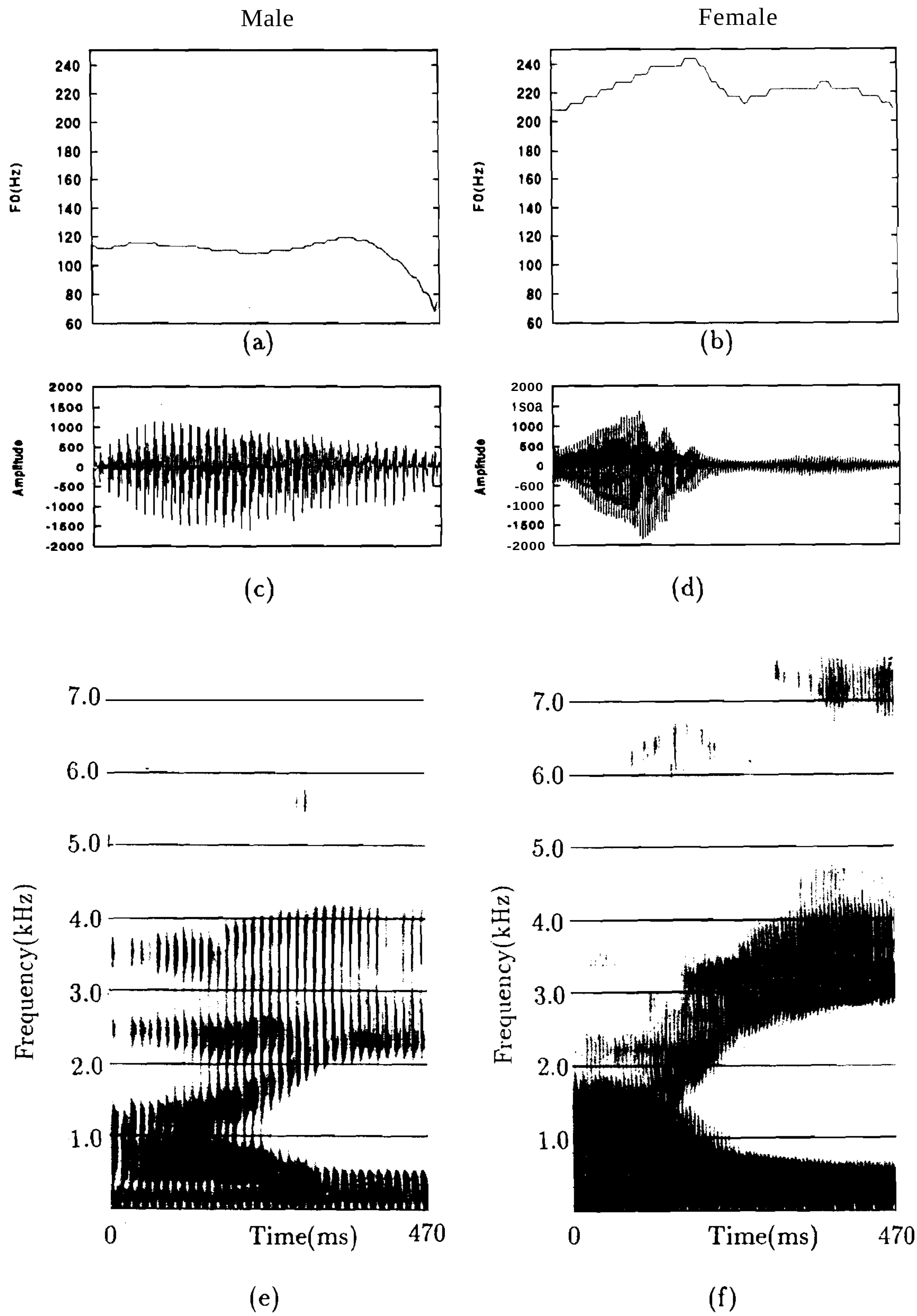


Figure 1.4: Illustration of the interspeaker variation at the segmental level. (c) and (d) shows the speech waveform for the sound /āī/. (a) and (b) show the pitch contours extracted from these utterances. (e) and (f) show the corresponding spectrograms.

which will transform the formant frequencies extracted from the speech of the source speaker to match with that of the target speaker. This transformation is to be captured from a limited amount of formant data and at the same time the captured transformation must be able to transform formants extracted from any utterance of the source speaker. Moreover such a transformation must not introduce distortions to smooth formant transitions occurring in continuous speech.

In order to capture the vocal tract system transformation between two speakers the vocal tract must he represented in a suitable manner. At one extreme the vocal tract can be represented by the envelope of the short time spectrum. But this representation is not motivated by the mechanism of speech production. Moreover, the short-time spectrum contains information related to both the vocal tract system and the voice source. The vocal tract system may be characterized by a linear time varying system represented by a set of time varying parameters. From a parameter extraction point of view, it is convenient to represent the system as a linear digital filter, for example an all-pole model. This representation takes into account the speech production mechanism upto some extent by modeling the vocal tract system as an all-pole filter. However, from a transformation point of view, it is desirable to represent the system with articulatory parameters. But articulatory parameters are difficult to extract from speech signal. Hence as a compromise, formants are proposed for representing the vocal tract system information. Formants are the resonances of the vocal tract system and thus they are very close to the physiology of speech production. At the same time in comparison with the articulatory parameters they are easy to extract from the speech signal.

In the context of speech perception the voiced segments, especially vowels, carry more information related to the speaker than consonants [9, 10]. There are two major reasons for this. They are: (i) Vowels are spectrally well defined and thus carry significant information about the vocal tract shape [3]. Since the vocal tract shape vary across speakers we can conclude that vowels and vowel like sounds (semivowels and dipthongs) carry important speaker specific information. (ii) Consonants are dynamic in nature and their durations are less in comparison with the durations of vowels. Hence while perceiving a consonant, the listener pays more attention in comprehending the message, i.e, recognizing the consonant. This causes the listener to ignore the speaker characteristics embedded in a consonant. In the case of vowels, since the duration is relatively large, the listener can pay attention to the voice characteristics also. This argument need not be valid for consonants like laterals and nasals which carry significant speaker specific information. But in these cases it is difficult to extract the speaker dependent parameters (for example, estimation of nasal resonant frequency). Hence in our studies we will be considering only voiced regions, especially vowels, for extracting the speaker characteristics.

It must be noted that even though gross speaker characteristics are attributed to the segmental factors, the real voice characteristics of a speaker are due to the manner of production acquired by the speaker over years. These learned factors may be present either through out an utterance (gross prosodic features) or only in some specific segments (segment specific prosodic features! of an utterance. Hence for voice transformation, both the gross and the segment based prosodic characteristics of the target speaker have to be incorporated into the synthesised speech, in addition to the segmental speaker characteristics. Performance of a voice conversion system critically relies on how well features that reflect the speech production mechanism can be extracted from speech. The quality of the transformed speech also depends on the synthesis scheme. In this work we use standard parameter extraction and synthesis procedures for voice conversion.

We first attempt to obtain transformation between vocal tract system of the source and target speakers using a single linear transformation. The linear transformation is derived using formants extracted from isolated utterances of vowels. The error in the transformed formants can be reduced significantly by using piecewise linear transformations. But piecewise linear transformations are capable of transforming formants extracted from steady vowels only. This transformation will introduce discontinuities in formant transition regions. These studies on linear formant transformation show that the formant scaling between two speakers is highly nonlinear. Since a feeclforward neural network with nonlinear computing elements is capable of capturing any arbitrary functional relationship, we used such a network to capture the inherently nonlinear formant transformation function. The network can be trained using formants extracted from isolated utterances of vowels. Even though such a network is capable of transforming formant transitions without introducing discontinuities, the transformed formant transitions were not as smooth as the target transitions. This failure of the network in transforming formant transitions properly is due to lack of the generalization capability of the network. We circumvent this problem by using appropriate training data to train the neural network. In this context we propose a method to test the generalization capability of the network by using synthetic test patterns. We also present a method by which the trained neural network can be efficiently used for voice conversion. This is done by using the network to modify the linear predictive coefficients (LPCs) extracted from the source speech and using the modified LPCs for synthesizing the transformed speech. Finally, we test this scheme of voice conversion by performing transformation between different speakers.

1.4 Review of related work

1.4.1 Introduction

The first attempt in voice transformation was reported in the classical paper by Atal and Hanauer [11]. In their paper Atal and Hanauer described the application of the LPC - vococler in modifying voice characteristics. In an experiment speech uttered by a male speaker was analysed to extract pitch, forinants and bandwidths. These parameters were modified using fixed scale factors. Speech was synthesized using these modified parameters to simulate a female voice. Seneff [12] deinonstrated a method by which the spectrum, the speaking rate and the pitch of speech signals could be modified even without extracting pitch. Even though this was a new speech analysis/synthesis system which was capable of independent manipulation of F_0 and the spectral enveiope, no study was carried out on its application to voice conversion.

In the above described efforts the main aim of the authors was not to convert the voice characteristics of a speaker to sound like that of another. They discuss voice conversion as an application of new methods of speech processing. The work reported by Childers et al (1985)[13] can be considered as one of the first attempts in voice conversion, since they were the first to focus on the problem of voice conversion in its own right. The following sections summarize studies made on voice conversion.

1.4.2 Voice conversion in a simple LP-synthesis framework

In this method, [14, 13] analysis was carried out on sentences uttered by the source and the target speakers to extract the speaker dependent information. From electro gloto graph(EGG) measurements, the average values of parameters T_1 and T_2, corresponding to the Fant's model [15] were measured, for different segments of the utterances of both the source and the target speakers. The values of the first three formants for different segments of the utterance of the source and the target speakers were determined. From this the scale factors for the three formants were computed. From the average pitch of the source and the target speakers, an average pitch scale factor was computecl.

In the transformation phase parameters extracted from speech of the source speaker were modified to correspond to the target speaker. The pitch was modified by the average pitch scale factor and the pitch contour was edited to match with that of the target pitch contour. The linear predictive coefficient(LPC) polynomial was solved to get the LPC roots. The roots which correspond to the first three formants were shifted in the z-plane in accordance with the scale factors computed during the analysis phase. The LPCs were recomputed from the modified roots. Then speech in the voice of the target speaker was synthesized using the modified LPCs, average pitch and pitch contour. To excite the LPC - vocoder, Fant's model was used during voiced segments and random noise for unvoiced segments. While using Fant's model, those model parameters (T_1 and T_2) which were measured during the analysis phase from the speech of the target speaker were used. According to the authors this method produced speech of good quality and the transformed speech possessed the speaker characteristics of the target speaker.

In a similar work done by Slifka and Anderson (1995), the scale factors for modifying the LPC roots were computed statistically. But the authors have reported that this method was not suitable for transforming the dynamic characteristics of the vocal tract [16].

1.4.3 Voice conversion by vector quantization

The method for voice conversion proposed by Abe et al [17] considers pitch, energy and spectral parameters as speaker dependent features. Spectral parameters were extracted from the utterances of the source and the target speakers and vector quantized. Similarly the extracted pitch values were scalar quantized. The correspondence between frames of the words uttered by the source and the target speakers were established by dynamic time warping (DTW) algorithm. This correspondence between the vectors of the source and the target speakers was accumulated as histograms. The histogram was used to represent each vector in the source speaker's codebook as a linear weighted sum of the vectors in the target speaker's code book. This correspondence is termed as the mapping code book. In the case of pitch frequency and gain, scalar quantization was used and the mapping codebooks for these parameters were defined based on the maximum occurrence in the histogram.

In the transformation phase, the speech of the source speaker was analysed to extract the speaker dependent parameters and were vector quantized using the source speaker's code hook. Using the mapping code book the corresponding vectors in the target speaker's code book were determined. Speech was synthesized using these parameter vectors.

In a similar work done by Savic and Nam (1991) the mapping code book was realized by a neural network [18].

In a later work done by Mizuno and Abe (1994), formant frequency modification was done by piecewise linear transformation rules to achieve voice personality transformation [19, 20]. The basic methodology of this technique is same as that suggested by Abe et al [17] except for the following points: (i) Instead of using LPCs formants were used. (ii) Spectral tilt was also considered for conversion. (ii) Instead of a mapping codehook, piecewise linear formant transformation rules were used to transform the formant frequencies and the spectral tilt.

1.4.4 Cross – language voice conversion

The objective in this work was to preserve the voice characteristics, when speech is translated from one language to another language and synthesised in the target language [21]. Hence the aim was to preserve the source speaker's characteristics in the synthesised speech across languages. The authors call this effort as cross – language voice conversion.

The major issue, related to the manipulation of speaker characteristics, lies in the incorporation of speaker characteristics into the speech synthesis system, which was used to synthesize the translated speech. The authors attempted a translation from Japanese to English. The translated text was synthesised using the MItalk system. The aim was to modify the output speech of the MItalk system so that the speech sounds like that of the Japanese speaker. To accomplish this voice transforrnation they used the mapping code book technique [17]. To build a mapping codehook, two speakers have to utter a set of training words. In this case the target speaker was a Japanese and the MItalk system represented the source speaker.

The converted speech was reported to be as intelligible as the MItalk output. In the case were the Japanese speaker was a female, the translated speech was judged by listeners as female speech.

1.4.5 Segment based voice conversion

Abe (1992) described a voice conversion system that used speech segments as the conversion units [22]. The speech of the source speaker was given to a speech recognition system for segmentation and labelling. To produce speech in the target voice, speech segments identified by the speech recognition system was replaced by the speech segments uttered by the target speaker. This system has a drawback that it depends on a speech recognition system for its performance.

1.4.6 PSOLA-based voice conversion

In this method proposed by Valbret et al (1992), the classical source-system decomposition was exploited to perform prosodic and spectral transformations [23, 24]. Prosodic modifications were applied on the excitation signal using TD-PSOLA [25] technique. The converted speech was then synthesized using the transformed spectral parameters.

For the spectral transformation, sentences uttered by the source and target speakers were time aligned by DTW. This procedure defines a mapping between the acoustic spaces of the two speakers. From this mapping the required spectral transformation was learned by first partitioning the acoustic space of the reference speaker by means of vector quantization (VQ) and by approximating the transformation within each class. The transformations associated with different classes were modeled in the training phase. Two methods were investigated for learning such a transformation, namely, Linear Multivariate Regression(LMR) and Dynamic Frequency Warping(DFW).

During the transformation phase, cepstral coefficients were extracted from each of the analysis frames of the input speech. The class to which the cepstral vector belongs was then identified by finding the nearest code vector. Then the transform related to this class was applied to the cepstral vector. This can be either the linear transformation(captured by the LMR technique) or the warping function (captured by the DFW technique). An LPC parameter set was extracted from the transformed cepstral or spectral vector, which was used in the synthesis of speech to reflect the voice characteristics of the target speaker.

This method worked well for short words. But in the case of sentences, due to differences in pronouncing, time alignments were imprecise and thus it was reported that the quality of the spectral transformation was degraded.

1.4.7 Comments

From the discussion of different approaches for voice conversion, it can be observed that none of them has addressed the issues related to voice characteristics in detail. All these methods were based on the use of existing signal processing techniques (like VQ, DTW, LMR) for performing voice conversion.

Most of the voice conversion algorithms found in the literature [17, 14, 21, 18, 22, 23, 19] depend critically on vector quantization of spectral parameters. Thus the source speaker's acoustic space was divided into separate nonoverlapping regions and for each such region a transform is statistically estimated. This transform is used to modify source spectral vectors belonging to that region, to match with that of the target speaker's spectral characteristics. It must be noted that spectral features directly correspond to the vocal tract characteristics. Hence separating the spectral space of the source speaker into discrete areas and transforming vectors of these areas separately would introduce discontinuities to the transformed vocal tract shape contour. Therefore the transformed spectral parameters may represent a discontinuous movement of the vocal tract and hence speech synthesized from such a set of parameters will he poor in quality. One solution is to increase the number of code vectors, which is equivalent to separating the spectral space rnore finely. Such an approach has the following disadvantages.

1. Increasing the number of code vectors would cause an increase in the storage requirements.
2. It will also increase the time required to search through a code book.
3. Most importantly, an increase in the number of code vectors will decrease the number of spectral vectors from which the transformation related to each of the distinct spectral spaces is estimated. Since this computation is done statistically, the estimate of the transform will become poorer as you separate the source speaker's spectral space into finer and finer regions.

Hence it is clear that better methods for transforming spectral vectors (vocal tract system information) are needed. All the voice conversion methods discussed above were successfully applied only to words or syllables. The issues in transforming the complex dynamics of the vocal tract system characteristics were left unaddressed.

The work reported by Childers et al [14] provides a good model for synthesizing speech from a set of acoustic parameters. In this method these acoustic parameters were extracted from the source speaker and transformed. Speech is synthesized using the transformed parameters. But the problem of learning these transformations were not addressed. In this thesis we mainly concentrate in the development of a neural network based system which could learn this transformation functions automatically. Emphasis is given to issues related to faithful transformation of the the dynamic features of the vocal tract system.

1.5 Organization of the thesis

The thesis is organized as follows. Chapter 2 cliscusses methods to capture linear functions which could transform the steady vocal tract system characteristics of the source speaker to that of the target speaker. This chapter also points out the advantages and drawbacks in approximating formant transformation by linear scale factors. The next chapter begins with an emphasis on the need for nonlinear approximation of formant transformation, and goes on to discuss methods for learning a nonlinear formant transformation using feedforward neural networks. Chapter 4 extends the application of feedforward neural networks for transforming formants extracted from dynamic speech sounds. Chapter **5** deals with incorporation of formant transformation into a voice conversion system and presents results of voice conversion between voices. Chapter 6 summarizes the work.

Chapter 2

Linear approximation of formant transformation

2.1 Introduction

This chapter discusses methods to capture the relation between formants derived from the speech of two speakers using linear transformation. We consider only the transformation of steady vocal tract shapes. For this purpose, formants are extracted from isolated utterances of steady vowels of the source and target speakers. We also describe a method for modifying the average pitch using a simple linear transformation. We first assume that the scaling corresponding to the vocal tract dimensions (mainly the effective vocal tract length) between speakers is linear, and hence the scaling of formants will also he linear. A simple linear formant transformation is captured using the least mean square algorithm (LMS). The transformation is same for all the formants. Error analysis is made to show that such a simple linear transform cannot efficiently transform the formants. We describe experiments to study the performance of using separate linear transformation for each formant. The error performance of the formant-dependent linear transformation is significantly better than the simple linear formant transformation. The shape arid the effective length of the vocal tract system is different for different vowels. The five vowels, /ā/, /ē/, /ī/, /ō/ and /ū/ are considered in this study. The vocal tract system transformation can be represented more efficiently by capturing linear transformations separately for each of the five vowels. Thus the entire formant transformation is represented by fifteen different linear transformations (3 formants x 5 vowels). This representation of the formant transformation, which is piecewise linear, is based on the assumption that the scaling of the vocal tract dimensions corresponding to the different vocal tract shapes(vowels) are nonuniform. A better approximation is obtained by using piecewise linear transformation instead of a simple linear transformation for each formant. In the case of piecewise linear transformation we need to know a priori the vowel from which the formant is extracted before applying the appropriate transformation. This is one of the drawbacks of piecewise linear transformation. Study of transformation of formant transitions show that the transformed formant transitions were not continuous. That is, piecewise linear transformations introduce discontinuities while transforming smooth formant transitions as in vowel sequences. These discontinuities are perceived during listening of the transformed speech.

The chapter is organized as follows. The next section gives a brief introduction to the LMS algorithm used for linear functional approximation. Section 2.3 describes the studies conducted in approximating the formant transformation using simple linear transformations. Experiments in capturing and testing the piecewise linear formant transformation is described in Section 2.4. Section 2.5 discusses a linear transformation to modify the average pitch of the source speaker to match with that of the target speaker. Synthesis experiments conducted to evaluate the performance of various linear and piecewise linear formant transformation schemes are discussecl in Section 2.6. The drawbacks in approximating the inherently nonlinear formant transformation by linear functions are discussed in Section 2.7.

2.2 Basics of linear functional approximation

The problem of capturing a linear rnapping function which maps the speaker dependent parameters extracted from the speech of the source speaker to those of the target speaker is formulated as follows:

Let the parameters extracted from the speech of the source and target speakers be represented by s^n, for n = *1, 2* ... N and t^n, for n = 1, 2 ... N respectively. It must he noted that $s^{\prime}$ and t^{J} are extracted from the corresponding speech segments uttered by the source and the target speakers. The problem is to find a linear function in the following form.

$$y^n = k_0 + k_1 s^n \tag{2.1}$$

The linear transformation given by equation *(2.1)* can also be implemented as a linear network. The objective is to compute the weights of this linear network (k_0 and k_1) such that the squared error given by

$$J = \frac{1}{2}E(t^n - y^n)^2$$

is minimum. In the above equation E denotes expectation and t^n represents the desired output. The error J as a function of k_0 and k_1 is called the error surface. It can be shown that the shape of this error surface is a paraboloid. Getting a least square solution involves finding the global minimum of the error surface.

Thesolution can be obtained iterativelyfrom any initial random weight setting. This can be clone by the classical LMS algorithm [26]. The weight updation is given by

$$k_i^{n+1} = k_i^n - \eta\nabla_i, \tag{2.2}$$

where

$$\nabla_i = \frac{\partial J}{\partial k_i}. \tag{2.3}$$

```
I. Initialization
    $k_0 = 0$
    $k_1 = 0$
for iteration number i=1,2 ...
    do n=1,2 ... N
        II. Filtering
            Compute
                $y^n = k_0 + k_1 s_1^n$
                $e^n = t^n - y^n$
        III. Weight adjustment
            Compute
                $k_0^{n+1} = k_0^n + \eta\ (t^n - y^n)$
                $k_1^{n+1} = k_1^n + \eta\ (t^n - y^n)s_1^n$
    enddo
until weights converge
```

Figure 2.1: Algorithm for computing a linear transformation function using the LMS algorithm.

In equation 2.2 the superscript represents the iteration number and η represents the learning rate which is a small constant. The weight is adjusted in the direction of the instantaneous gradient of error. From equation 2.2 and 2.3 the weight adjustment becomes

$$k_i^{n+1} = k_i^n + \eta(t^n - y^n)s_i^n. \tag{2.4}$$

The LMS algorithm, for converting the source parameters to target parameters is given in Figure 2.1.

2.3 Learning linear formant transformations

Data for this study consists of isolated utterances of vowels /ī/, /ē/, /a/. /ō/, and /ū/ from five male and five female speakers. Each of these vowels was repeated by every speaker twenty times. Out of these twenty sets of vowel data, fifteen sets were labeled as the training set and the remaining were considered as the test set. The first three formants were extracted using a method based on minimum phase group delay functions [27].

In the first attempt. a simple linear network was trained using the LMS algorithm to transform all the formants in the same manner. A male speaker was considered as the source speaker and a female speaker as the target speaker. The error between the source and the target formants before and after the application of the linear transformation is shown in Table 2.1. Let the source, target and the transformed formants be denoted by by F_s^i, F_t^i and F_{tr}^i for i = 1, 2 ... N, respectively. N is the number of frames from which the formant data was extracted. The error in percentage between the the target and the source formants is calculated as $^{1}\sum_{i=1}^{N}\frac{|F_t^i-F_s^i|}{F_s^i}$. Similarly the error in percentage between the target and the transformed formants is given by $\frac{1}{N}\sum_{i=1}^{N}\frac{|F_{tr}^i-F_t^i|}{F_t^i}$. From the table we can note that even though the application of the linear transformation brings the source formants closer to the target formants, the error is still large. This is evident in the case of F_1 for the vowel /ē/ and F_2 for the vowel /a/. Also note that in some cases (F_2 of vowel /ō/ and F_1 of vowel /ē/), the error between the source and the target formant is less than the error between the transformed and the target formant. It means that in these cases the application of the transformation to the source formants moves it farther away from the target formant location.

For reducing the error in the transformation, three separate linear networks were trained to capture transformations corresponding to each of the three formants. Thus the formant transformation now consists of a set of three linear functions. Figure 2.2 shows the three learned functions. They are represented as straight lines.

Table 2.1: Error analysis for the single linear formant transformation.

vowels	Error in percentage between the target and the source'formants			Error in percentage between the target and the transformed formants		
	F_1	F_2	F_3	F_1	F_2	F_3
/ī/	**27.1**	**14.3**	**4.8**	9.0	**3.6**	**6.8**
/ē/	7.0	32.9	18.8	19.8	16.2	5.4
/ā/	42.0	21.0	10.1	25.3	18.0	3.8
/ō/	7.3	4.9	5.4	9.9	12.0	5.6
/ū/	21.5	4.2	9.3	14.4	11.2	3.7

Note that the three lines shown in the Figure 2.2 have different slopes. This shows that the scaling factors for the three formants are different even for steady vowels. These functions were used to transform the formants of the test set. Table 2.2 gives the errors between the source formants and the target formants before and after the application of the three linear transforms. By comparing Table 2.1 and Table 2.2 it is clear that by using separate linear networks for the three formants we have been able to reduce the error in the transformed formants significantly. This improvement is significant in the case of second formant (compare the sixth columns of Table 2.1 and Table 2.2). For some cases the formant-dependent transforniation gives larger error in comparison with the case of a single linear transformation (for example, F_1 of vowel /ī/ and F_3 of vowel /ū/). But if we consider the overall error, the use of formant-dependent transformation outperforms the single linear transformation.

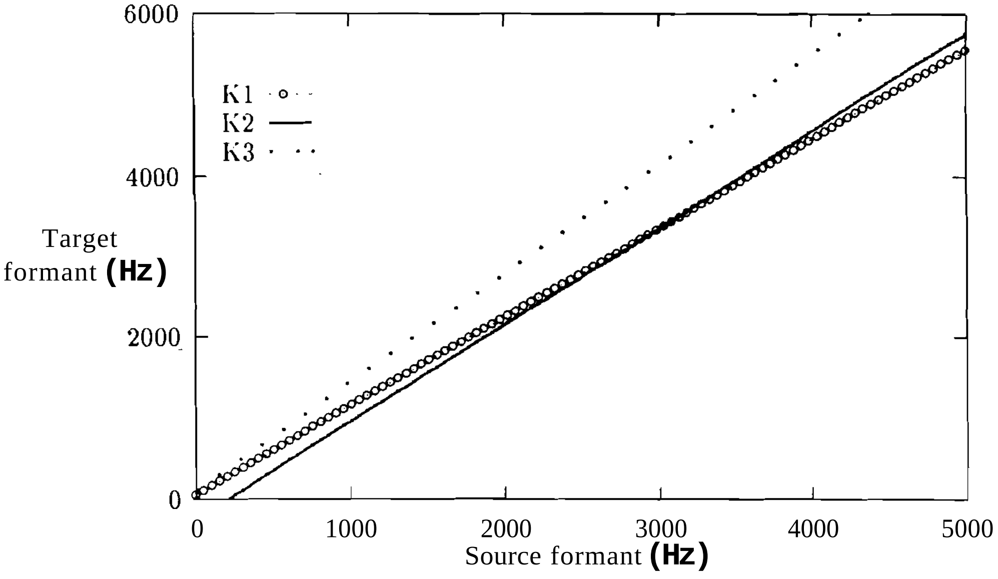


Figure 2.2: The formant dependent transforrnation function. Figure showing the transformation functions corresponding to the first (kl), second(k2) and the third(k3) formants.

2.4 Piecewise linear formant transformation

The vocal tract shape and effective length are significantly different for different vowels. Therefore the vocal tract system transformation can be improved significantly by using separate linear transformations to transform formants corresponding to different vowels. This is equivalent to making the transformation of each formant piecewise linear. For five vowels and three formants, the number of linear functions will be fifteen. Thus we have trained fifteen linear networks to capture the formant transformations for all the five vowels. Each of these fifteen networks is expected to transform a formant (first, second or third) extracted from one of the five prototype vowels (/ā/, /ē/, /ī/, /ō/ or /ū/). Figure **2.3** shows the resulting piecewise linear transformation function. The figure shows the scale factors of the formants for five prototype vowels. Scale factors refer to the amount by which the source formants corresponding to different prototype vowels are scaled by the transformation. In

Table 2.2: Error analysis for the formant dependent linear transformation.

Vowels	Error in percentage between the target and the source formants			Error in percentage between the target and the transformed formants		
	F_1	F_2	F_3	F_1	F_2	F_3
/ī/	27.1	14.3	4.8	12.0	4.8	5.1
/ē/	7.0	32.9	18.8	16.2	6.8	4.6
/ā/	42.0	21.0	10.1	21.0	4.5	4.7
/ō/	7.3	4.9	5.4	7.0	3.2	4.2
/ū/	21.5	4.2	9.3	11.0	2.6	3.9

this study we have considered five different sets of source and target speakers. The scale factors shown in the figure correspond to transformation of formants of a male speaker to those of a female speaker. These results show that the scale factors are dependent both on the formant (first, second or third) and the quality of the vowel. The variations of the three scale factors show similar trend across different sets of male and female speakers. A notable deviation from the uniform scaling of the formants is the large scale factor for the first formant corresponding to the open vowel /ā/ in comparison with the closed vowels /ū/ and /ī/. The scale factor for the second formant is high for front vowels /ī/ and /ē/. Note that for the back vowels /ū/ and /ō/ the value of the second formant scale factor is less than unity. This means that the second formant frequency for back vowels /ū/ and /ō/ is higher for male speakers than for the female speakers. These observations are consistent with a similar study conducted by Fant [28].

In order to transform a set of formants using piecewise linear transform, it is necessary to first identify the vowel from which the formants were extracted. We

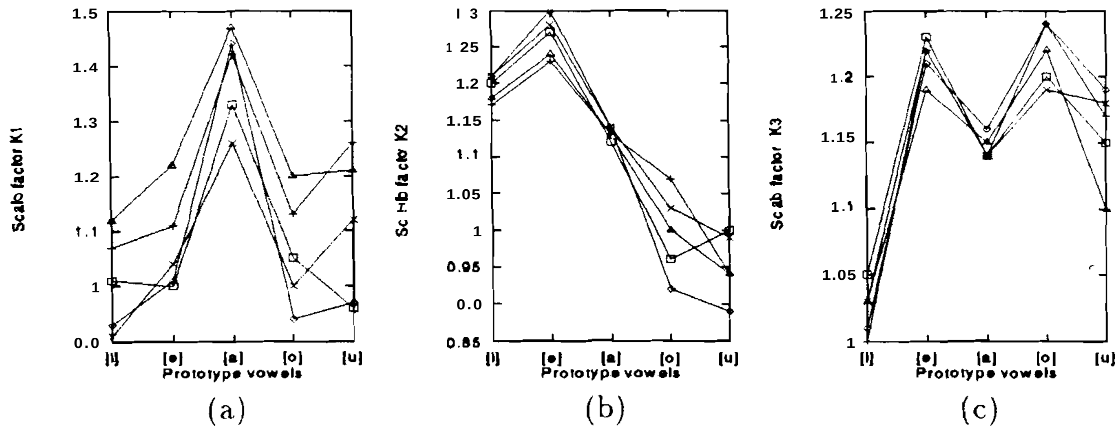


Figure '2.3: The piecewise linear formant transformation function (a) Scale factor for the first formant (b) Scale factor for the second formant (c) Scale factor for the third formant. Each set corresponds to one pair (male-female) of speaker data.

have used a simple classification scheme for recognizing the vowel from the formants. Figure 2.4 shows the classification scheme used for transforming the source formants using the piecewise linear transformation. We first compute the mean of the formant values for different vowels. The formant vectors $\bar{F}_i$, $\bar{F}_e$, $\bar{F}_a$, $\bar{F}_o$, and $\bar{F}_u$ represent the mean of formant vectors corresponding to the vowels /ī/, /ē/, /ā/, /ō/, and /ū/. Now given a formant vector $\bar{F}$, we compute the Euclidean distance between this unknown formant vector and the mean formant vectors corresponding to the prototype vowels. We recognize the unknown speech sound as the vowel whose mean formant vector is closest to the formant vector extracted from the unknown speech sound. Once speech sound is recognized as one of the vowels, the appropriate linear transformation is applied to each of the formants. Table 2.3 gives the error analysis of the piecewise linear transformation. From Tables 2.1, 2.2 and 2.3, it can be observed that when the formant transformation function is approximated by a single linear function the error during testing is the highest. When we use three separate linear functions for the three formants, the error reduces. The error was

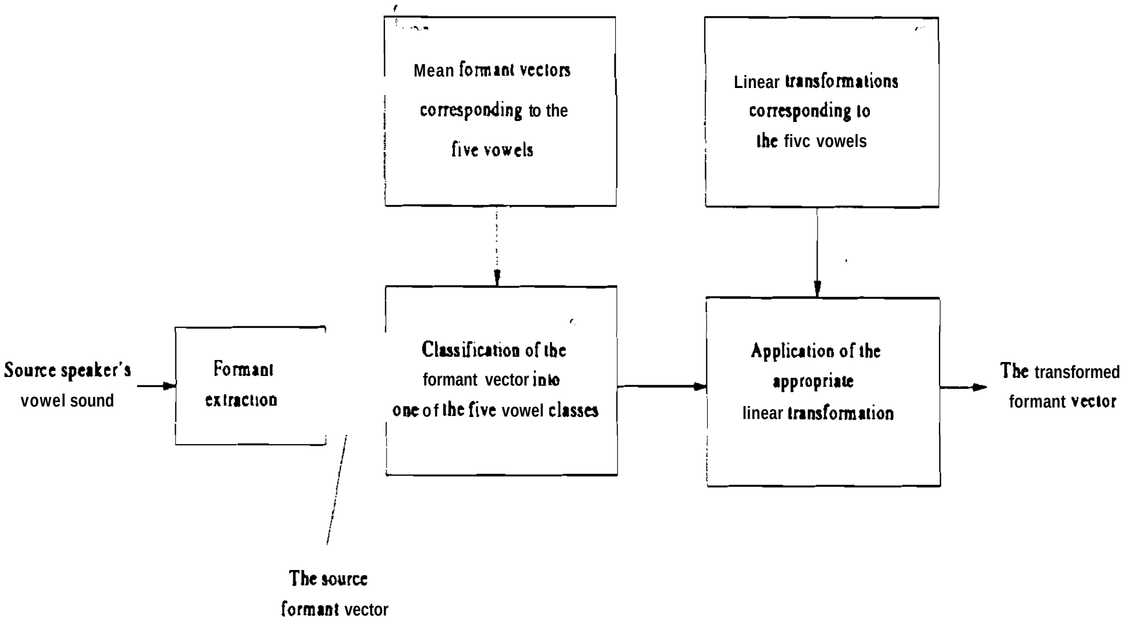

Mean formant vectors corresponding to the five vowels
Linear transformations corresponding to the five vowels
Source speaker's vowel sound
Formant extraction
Classification of the formant vector into one of the five vowel classes
Application of the appropriate linear transformation
The transformed formant vector
The source formant vector

Figure 2.4: Block diagram showing the transformation of the formants extracted from vowels using the piecewise linear transformation.

found to be least when these functions themselves were made piecewise linear.

2.5 Pitch transformation

Average pitch also contributes to the voice characteristics and thus for synthesizing speech in the voice of the target speaker, the average pitch of the source speaker has to be modified. In this section we describe a simple procedure to modify the average pitch of the source speaker, using linear networks.

The task is to find a linear transformation using a linear network which could modify the average pitch extracted from the speech of the source speaker to match with that of the target speaker. For this, pitch was extracted from the training set of isolated utterances of vowels. For extracting pitch we have used the SIFT algorithm [29]. The F_0 data correspontling to the source and the target speakers is

Table 2.3: Error analysis for the piecewise linear formant transformation.

vowels	Error in percentage between the target and the source formants			Error in percentage between the target and the transformed formants		
	F_1	F_2	F_3	F_1	F_2	F_3
/ī/	27.1	14.3	4.8	3.8	1.7	1.7
/ē/	7.0	32.9	18.8	4.2	2.3	3.2
/ā/	42.0	21.0	10.1	9.2	1.2	3.4
/ō/	7.3	4.9	5.4	3.3	1.9	2.3
/ū/	21.5	4.2	9.3	4.8	2.5	1.9

represented by F_{0s}^{i}, for $i = 1, 2 \ldots$ N and F_{0t}^{i}, for $i = 1, 2 \ldots$ N respectively, where N is the total number of training values of F_0. The weights of the linear network were computed using the LMS algorithm described in Section 2.2. Once we have captured the linear pitch transform, the next issue is to validate the function for its capability to transform the average F_0. For evaluation, F_0 contours are extracted from the test utterances of both the source and the target speakers. The source pitch contours are transformed using the linear pitch transformation. Table 2.4 shows the error in the average F_0 of the source and the target pitch coritours before and after the application of the transformation function. From the table it is clear that the average pitch transformation is able to modify the average pitch of the source speaker to match satisfactorily with that of the target speaker.

Table 2.4: The error analysis on the linear F_0 transformation.

vowels	Error in percentage between the target and the source F_0	Error in percentage between the the transformed F_0 and the target F_0
/ī/	51.2	9.2
/ē/	45.4	7.3
/ā/	50.9	8.2
/ō/	40.4	7.1
/ū/	47.8	6.9

2.6 Synthesis experiments

Error analysis gives only a quantitative idea of the performance of the learned function. A more useful way of evaluating the transformation is to synthesize speech from the transformed formants and the transformed pitch. We have extracted formants, pitch and gain from a set of vowels uttered by the source speakers (male). The pitch was transformed by the learned linear function. The formants were modified using the three different methods of formant transformation scheme described in Section 2.3. The speech was synthesized with the modified average pitch and formants. A standard glottal source model was used to excite the formant vocoder. Informal listening of the transformed speech was used to judge the quality of the voice conversion. When a single linear function was used to transform the source formants, the voice characteristics in the converted speech was quite different from that of the target speaker. But by using three separate functions for modifying the three formants the quality of the voice conversion improved. When piecewise linear functions were used to modify the source formants, the voice characteristics of the converted speech were close to that of the target speaker.

2.7 Limitations of linear formant transformation

From the formant scale factors captured by the linear network it is clear that the scaling of the formants is dependent on the vocal tract shape. Even for transforming steady speech sounds like vowels the formant transformation is nonlinear. Hence we have used piecewise linear functions to approximate the inherent nonlinear formant transformation function. There are mainly two disadvantages in using a piecewise linear formant transformation. They are: (i) If we are using separate linear transforms for different vowels, we will have to identify the speech sound as one of the vowels before applying the appropriate transformation to the formants extracted from that speech segment. (ii) If we use a piecewise linear function to approximate the inherent nonlinearity in the transformation, the resulting formant transformation function will have discontinuities.

Figure 2.5 shows the effect of transforming various types of formant transitions using the piecewise linear formant transformation. The first column shows the formant contours extracted from the speech of the source speaker corresponding to the vowel sequences /āī/, /āū/ and /ōī/. The third column shows the corresponding target formant transitions. The source formant transitions were transformed using the piecewise linear transformation. The second column of the figure shows the transformed formant transitions. Note that the transformed formant transitions show discontinuities. These discontinuities are significant in the case of the third formant corresponding to the vowel sequences /āī/ and /ōī/. These discontinuities occur between the tenth and the eleventh frames. This is due to the use of two separate linear functions to transform the two different parts of the formant transition. It must be noted that such a discontinuous formant contour represents an abrupt change in the vocal tract system which is undesirable, as it degrades the quality of the synthetic speech generated from such discontinuous formant contours.

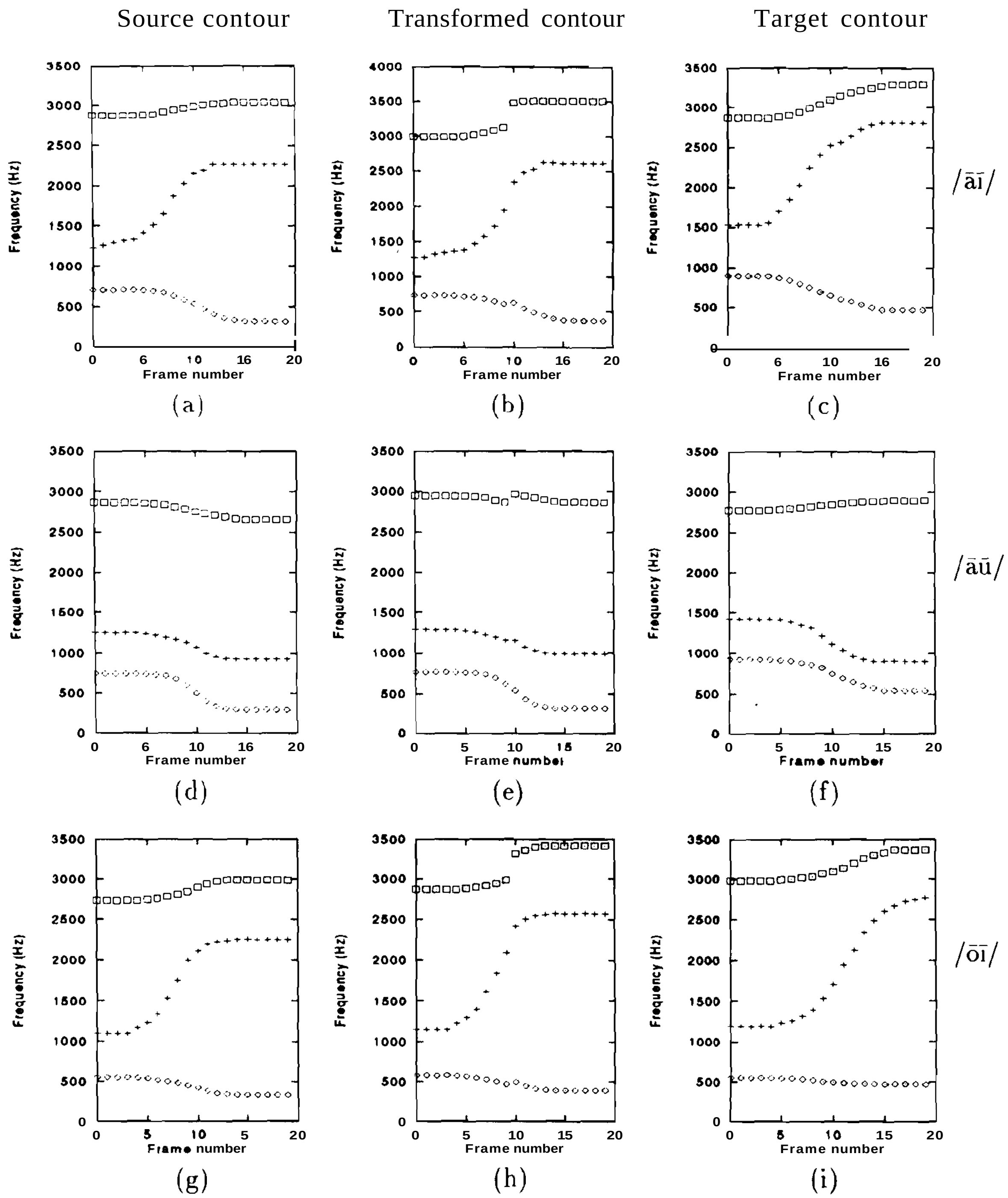


Figure 2.5: Figure illustrating the problem of formant discontinuity. (a),(d) and (g) show the source(male) formant contours corresponding to the vowel sequences /āī/, /āū/ and /ōī/. (b),(e) and (h)show the corresponding transformed formant contours using the piecewise linear transformation. (c), (f) and (i) show the corresponding target formant contours (female).

2.8 Summary

In this Chapter we have discussed methods to capture linear functions for transforming formants and average pitch of the source speaker. The major advantage of using a linear network to capture a mapping function is that, the convergence is guaranteed while searching the weight space for the optimum solution. The disadvantage is the large error in the transformed formants when we use a single function to approximate the required transform. If piecewise linear transform is used, the error in the transformed formants will be reduced. But we have to know a priori the speech sound from which the formants were extracted in order to apply the appropriate transformation. If we use such a piecewise linear transform to modify a formant transition, the transformed formant contour will have discontinuities. When this transformed formant contour is used for synthesizing speech, the synthetic speech will represent an abrupt change in the vocal tract movement and thus will be unnatural. In the next chapter we describe how a neural network can be trained to capture the inherently nonlinear transformation of the formants.

Chapter 3

Neural network for formant transformation,

3.1 Introduction

This chapter describes neural network models to capture the nonlinear formant transformation function. In the previous chapter we found that if we use a linear transformation on source formants, then the resulting error in the transformed formants is very high. This is because the transformation from male formants to female formants or vice versa is highly nonlinear. A rnultilayer feedforward neural network with nonlinear processing elements is capable of capturing any arbitrary mapping function [30]. Hence we propose to use such a network to capture the mapping function which transform the formants of the source speaker to those of the target speaker. We describe how such a network can be trained to capture the required formant transformation operation. Formants extracted from isolated utterances of vowels are used to train the network. In the case of neural network-based fbrmant transformation, the network can be used to transform a formant vector without knowing the class of the input. vector. This is an advantage in using a neural network in transforming formants. Such a network is useful for transforming formant

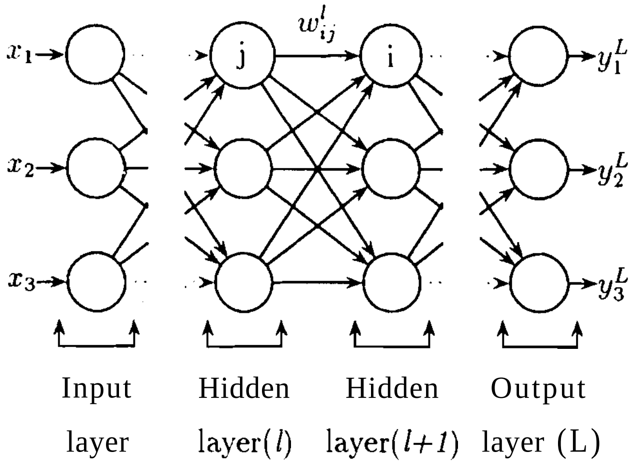


Figure 3.1: The architecture of a multilayer feedforward neural network.

transitions also. We will obtain smooth transformed formant transition as in natural speech although they may be significantly different from the target formant transition.

In Section 3.2 the theory of backpropagation (BP) algorithm is presented briefly, as the algorithm forms the basis for training the neural network. Section 3.3 discusses the studies conducted to capture the formant transformation by a neural network. Section 3.4 summarizes the results of this chapter.

3.2 Backpropagation(BP) algorithm

Figure 3.1 shows the structure of a feedforward neural network. The network shown in the figure has two hidden layers besides the input and output layers. The nodes in the input layer are linear, whereas the nodes in the hidden and output layers are nonlinear processing units. Figure 3.2 shows the nonlinear output (sigmoidal) function used in the nodes of the hidden and output layers. The output of each of the nodes is given to the input of each of the nodes in the next layer after linearly

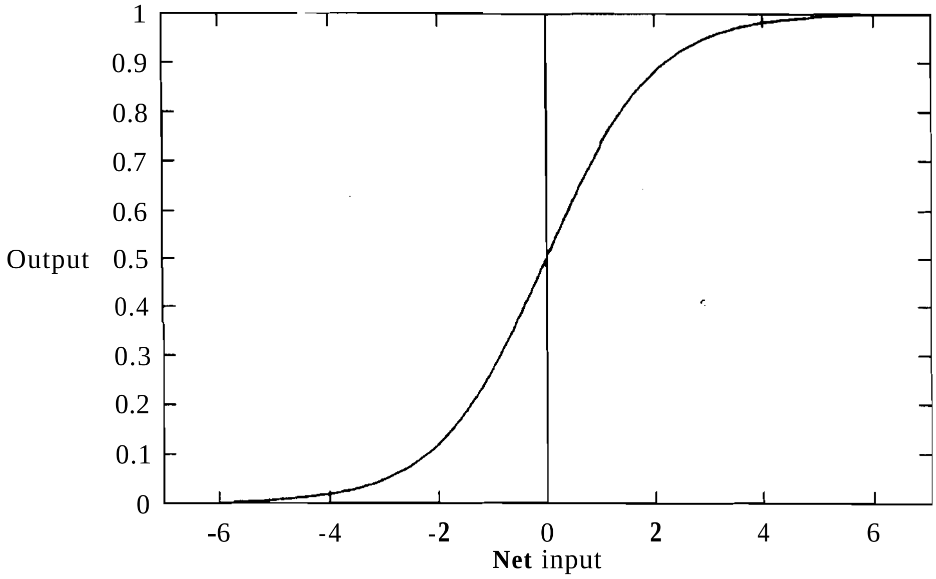


Figure 3.2: The sigmoidal nonlinearity used as output function in the nodes.

weighting it. The output of each of the nodes, other than the ones in the input layer, is obtained by adding up all the inputs and passing the sum through the sigmoidal nonlinearity shown in Figure 3.2.

It has been proved that a network as shown in Figure **3.1** is capable of representing any arbitrary function [30]. The next issue is how to capture the required input-output relation from a limited amount of data. Backpropagation algorithm or the generalized delta rule [31, 32] is an algorithm which can be used to adjust. the weights of the network, so that the network captures the implicit function represented by a set of input-output vectors.

The weights are initialized to randorn values. The first training input X^1 is given as input. The input, v_i^li, to the i^{th} node in the l^{th} layer is computed by the following equation.

$$v_i^l = \sum_j y_j^{l-1} w_{ij}^{l-1}$$

where y_j^{l-1} is the output of the j^{th} node in the $(1-1)^{th}$ layer arid w_{ij}^{l-1} is the weight connecting the i^{th} node in the l^{th} layer and the j^{th} node in the $(l-1)^{th}$ layer. The output of a node in the input layer is same as the input of the node. The output of the nodes in the hidden and the output layers are calculated using the following equation.

$$y_i^l = \varphi(v_i^l)$$

where $\varphi(.)$ is the nonlinear output function.
This forward computation will give the output, y_i^L for i=1,2 ... N, of the network. Here L is the number of layers and N is the number of nodes in the output layer. Then the error at the output layer is given by

$$e_i^L = d_i - y_i^L \qquad\qquad i = 1, 2, \ldots N.$$

The error e_i^L is used to adjust the weights. For adjusting weights, a weight correction $\Delta \mathrm{w}_{ij}$ is defined by the generalized delta rule, which is added to the weight w_{ij}:

$$\begin{aligned} (\textit{Weight correction}) \; &= \; (\textit{learning parameter}).(\textit{local gradient}). \\ & \quad\; (\textit{input to that weight}) \end{aligned}$$

which becomes

$$\Delta w_{ij}^l = \eta \delta_i^{l+1} y_j^I,$$

where η is the learning parameter and δ_i^{l+1} is the local gradient. The local gradient at any node **i** in the output layer is given by

$$\delta_i^L = e_i^L y_j^L (1 - y_j^L).$$

Then the weights are acljusted using the following formula.

$$w_{ij}^{L-1}(n+1) = w_{ij}^{L-1}(n) + \eta \delta_i^L y_j^{L-1} \tag{3.1}$$

The local gradient at any node i in the layer (hidden) l is given by

$$\delta_i^l = y_i^l (1 - y_i^l) \sum_k \delta_k^{l+1} w_{ki}^{l+1}.$$

then the weight adjustment becomes

$$w^l_{ij}(n+1) = w^l_{ij}(n) + \eta\delta^{l+1}_i y^l_j. \qquad (3.2)$$

Thus, the output weights and the hidden weights are adjusted after presenting the training patterns. If all the training vectors are presented to the network and the weights are adjusted as given in equations **3.1** and **3.2** , then we say that an epoch of training is over. Training will have to be done for several such epochs until the weights converge. Generally the back propagation algorithm is very slow in converging to a solution. By introducing a momentum term in the learning equation 3.2, the rate of convergence can be improved significantly. Then the equation **3.2** becomes

$$w^l_{ij}(n+1) = w^l_{ij}(n) + \eta\varphi^l_i y^l_j + \alpha[w^l_{ij}(n) - w^l_{ij}(n-1)], \qquad (3.3)$$

where a is a small positive constant.

3.3 Studies in formant transformation using the BP network

Figure **3.3** gives the block diagram which shows the computation of the required transformation function using a neural network trained using the back propagation algorithm. In the training phase we input formant values extracted from isolated utterances of vowels of the source speaker to the network. The desired output is the formant values extracted from the corresponding utterances of the target speaker. Data used for this study is same as that used in Section **2.3.** After training, it is expected that the neural network would have captured a function which maps the formants of the source speaker to that of the target speaker. A multilayer feed forward neural network with two hidden layers was used for this purpose. We have used three elements each in the input, hidden and output layers. The steps involvetl

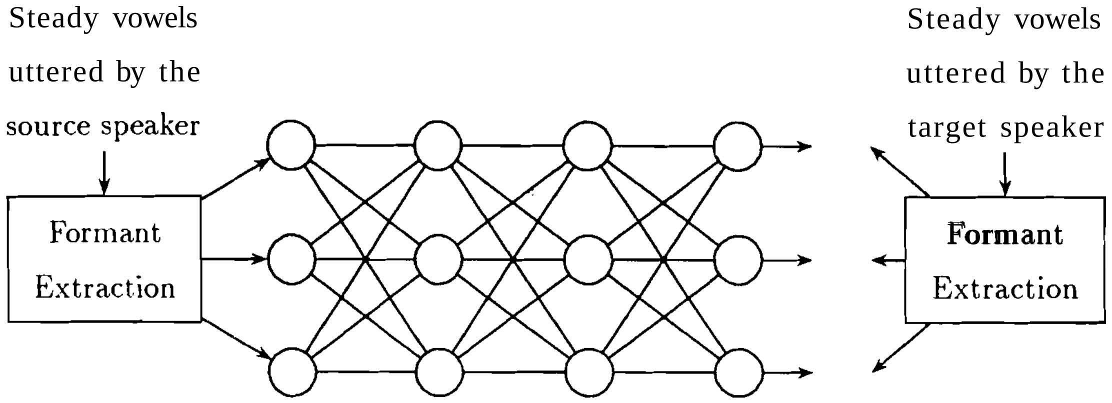

Steady vowels uttered by the source speaker
Formant Extraction
Steady vowels uttered by the target speaker
Formant Extraction

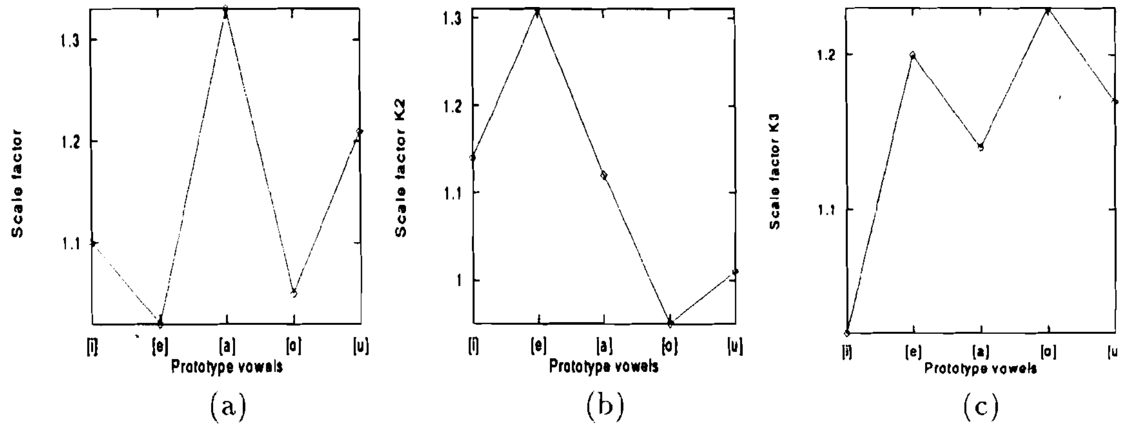
Figure 3.3: Training a neural network for capturing a mapping function which transforms the formants of the source speaker to that of the target speaker.

in training a neural network for capturing the formant transformation function are shown in Figure **3.4.** Figure 3.5 shows the transformation learned by the network. The transformation is shown in the form of scale factors by which the source formants corresponding to the various prototype vowels are scaled by the network. We have considered here a male-to-female formant transformation. Comparing the transformation learned by the neural network (Figure 3.5) and the piecewise linear transformation (Figure 2.3), we observe that the shape of these functions are nearly the same. This shows that formants extracted from steady vowels are transformed in a similar fashion hy both the piecewise linear transformation and the neural network. Hut it must he notecl that the way in which the transformation is carried out by the linear transformation and the neural network are significantly different. In the case of piecewise linear transform, the transformation is described by a set of fifteen

```
repeat
  For each set of formant data
      begin
         Step - I
            The formant values ($F_1$ to $F_3$) corresponding to
            the source speaker are given as input
         Step - II
            The formants extracted from the
            same vowel uttered by the target speaker are fed as
            the desired output
         Step - III
            The weights are adjusted using the backpropagation algorithm
      end
until weights converge
```

Figure **3.4:** Algorithm for training the feedforward network to capture formant transformation function.

simple linear transforms. In order to transform a formant vector we need to know the vowel class of the speech segment from which the formant vector is extracted. Moreover, the piecewise linear nature of the formant transformation function will introduce discontinuities while transforming smooth formant transitions. In the case of neural network the formant transformation is captured as a single continuous nonlinear fiinction. Hence it is not necessary to classify the input formant vector before transforming it. Since the transformation captured by the network is inherently nonlinear, the network will transform the formants appropriately depending upon the value of the formant. This avoids the necessity of knowing the class of the input formant vector. This is a significant advantage of using neural networks for capturing the inherently nonlinear formant transformations. This will be evident while transforming formant transitions.



(a)

(b)

(c)

Figure 3.5: The scale factors learned by the network. (a), (b) and (c) show the scale factors for first, second and third formants.

It is necessary to study how far the network is successful in capturing the relation between the formants of the source and the target speakers. For testing the network formants were extracted from test utterances of the source speakers (male). and were transformed using the trained neural network to get the transformed formants. We compared the error between (a) the target and the source formants and (b) between the target and the transformed formants. Table 3.1 shows the percentage error between the formants of the target speaker and the source speaker before and after the application of the transformation learned by the neural network. From this table it is clear that the application of the transformation learned by the neural network to the source speaker's formants has resulted in a significant reduction in the error between the transformed and the target formants.

We have also examined the capability of the network to transform formant transitions. For this we have extracted formants from vowel sequences, /āī/, /āū/ and /ōī/ for the source and the target speakers. The formant contours obtained from the speech of the source speaker was transformed using the trained network.

Table 3.1: Error analysis on the network which was trained using formants extracted from vowels uttered in isolation.

Vowels	Error in percentage between the target and the source formants			Error in percentage between the target and the transformed formants		
	F_1	F_2	F_3	F_1	F_2	F_3
/ī/	27.1	14.3	4.8	4.0	2.7	2.1
/ē/	7.0	32.9	18.8	1.0	2.4	2.3
/ā/	42.0	21.0	10.1	4.1	1.4	2.4
/ō/	7.3	4.9	5.4	3.0	2.1	1.3
/ū/	21.5	4.2	9.3	3.2	2.4	2.0

Figure 3.6 illustrates the problems of discontinuity in the transformed formant contours. The first column of the figure shows the source formant transitions. The second column shows the formant transitions obtained by transforming the source formant contour by using a trained neural network. The corresponding target formant transition is shown in the third column of the figure. From the figure it is clear that the network is capable of transforming formant transitions without introducing discontinuities. But we have noted that the transformed formant contour is not same as that of the target contour. The deviation is more pronoutrced in the case of the second formant corresponding to the sounds /āī/ and /ōī/.

3.4 Summary

In this chapter we have described a neural network model to capture the inherently nonlinear transformation of the formants across speakers. The backpropagation algorithm was used to train the network with the formants extracted from steady

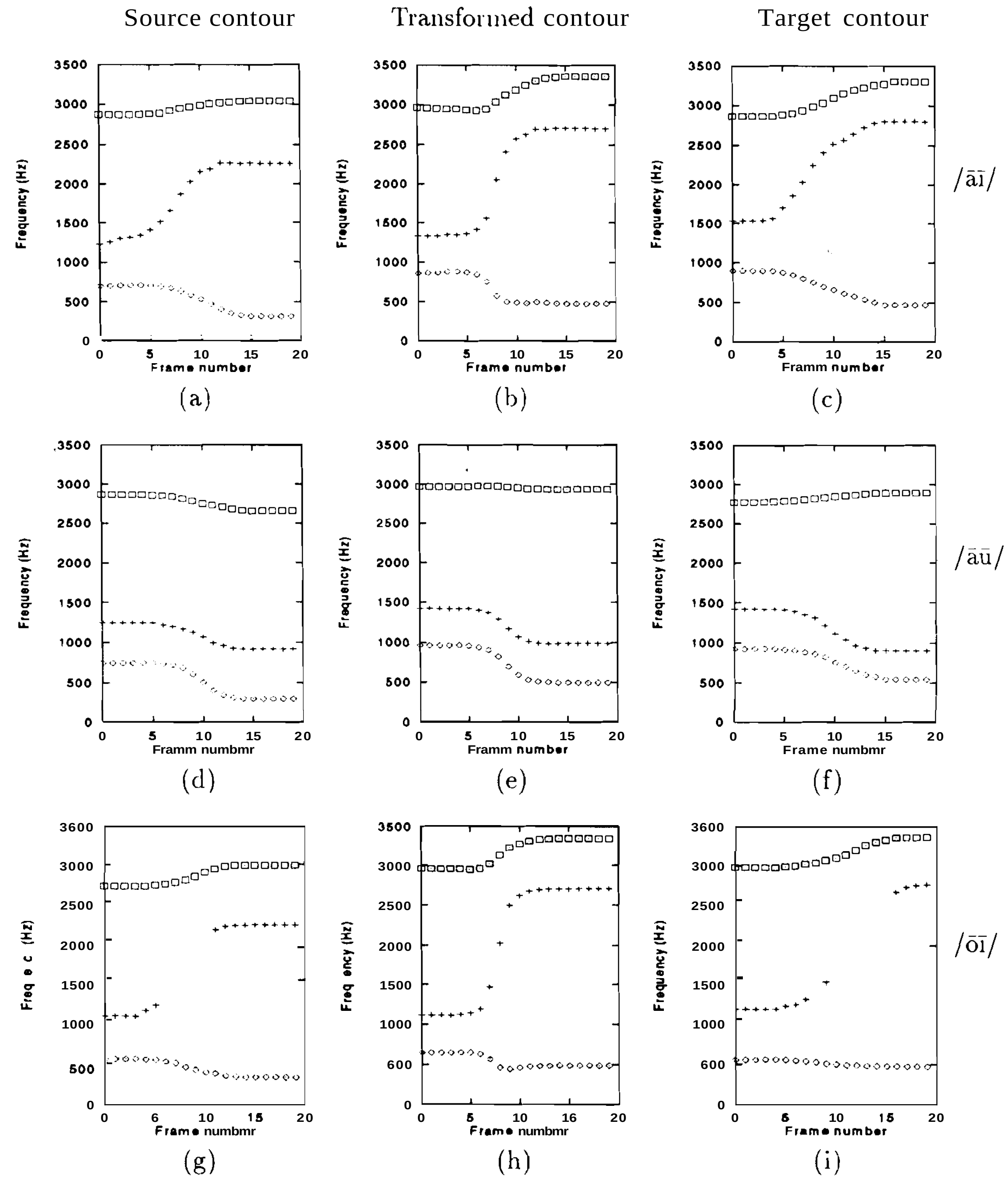


Figure 3.6: Problems of formant contour smoothness. (a),(d) and (g) show the source (male) formant contours corresponding to the vowel sequences /āī/, /āū/ and /ōī/. (b),(e) and (h) show the corresponding transformed formant contour using the trained network. (c), (f) and (i) show the corresponding target formant contours (female).

vowels. We have demonstrated that such a network is capable of overcoming some of the problems posecl by linear approximation of the formant transformation. Even though the trained network can transform formant transitions without introducing discontinuities into the formant contour, it was observed that the transformed formant transitions were steeper than the target formant transitions. Hence it becomes clear that a network trained using the formants extracted from steady vowel sounds is not capable of transforming formant transitions. In the next chapter we discuss methods to train a feedforward network to capture a transformation which can transforni not only steady formants but also formants extracted from dynamic speech sounds.

Chapter 4

Transformation of dynamic sounds

4.1 Introduction

In natural speech the vocal tract system continuously changes its shape. This is manifested as transitions in the formant contours extracted from continuous speech. In order to transform the formants extracted from continuous speech, it is necessary to transform the formant transitions as well. This chapter focuses on the problem of capturing the dynamic characteristics of the vocal tract system. In Chapter 3 we have observed that a neural network trained using the formants extracted from steady vowels was not able to transform formant transitions properly. Hence in this chapter we explore methods for transformation of the dynamic characteristics of the vocal tract system, which are manifested as formant transitions.

The failure to capture formant transitions by a network trained with formants data extracted from steady vowels is due to lack of *generalization* in the network. Generalization in a network refers to the ability of the network in giving correct outputs to inputs for which it was not trained. This lack of generalization capability in turn is clue to nonrepresentative nature of the training data, when collected from steady vowel regions. In this chapter we demonstrate that a BP network trained with representative data can transform not only steady formants but also formant transitions. We also address the issue of testing the generalization capability of such a network. It is not possible to test the performance of the network by using formant transitions because the target and the transformed formants cannot be compared directly due to warping in time. Moreover, there may be variations in the formant trajectories due to interspeaker variations and coarticulation. We propose a method using synthetic formant transition data to test the generalization capability of the network.

The following section gives a brief introduction to the notion of generalization. A method to improve the generalization is discussed in Section **4.3.** Section 4.4 discusses the proposed method for testing the generalization capability of a network in the context of transforming formant transitions.

4.2 The problem of generalization

The generalization capability of the network is mainly determined by the following four factors [32]:

1. *Training data:* This refers to how well the training data set represents the input-output mapping.

2. *Architecture of the network:* The ar hitecture refers mainly to the size of the network. If one uses a network size which is too large, it may lead to memorization of the examples used for training and thus will result in poor generalization.

3. *Training methodology:* The training algorithm also influences the generalization performance of a network. If we assume that the standard backpropagation algorithm is used for training, then the issue is when to stop the training. Overtraining of a network will cause overfitting of the training data and in turn will lead to poor generalization.

Table 3.1: Mean (M) and variance (V) of the formant frequencies extracted from steady vowels uttered in isolation by the source and the target speakers.

Vowels	Source speaker						Target speaker					
	F_1		F_2		F_3		F_1		F_2		F_3	
	M	V	M	V	M	V	M	V	M	V	M	V
/ā/ /ī/	728	16	1244	48	2773	17	922	32	1419	25	2895	48
/ē/	554	24	1919	24	2587	17	525	21	2540	61	3055	99
/ī/ /ā/	333	15	2245	37	2985	40	470	19	2722	37	3298	110
/ō/	549	9	1098	18	2738	21	592	22	1099	29	2893	63
/ū/	385	8	1059	51	2595	18	542	27	1079	28	2833	65

3. *The inherent complexity of the problem:* The complexity of the mapping function which one wants to capture also influences the generalization performance.

The failure of a network, trained using the formants extracted from steady vowels, in transforming formant transitions is considered as the lack of generalization capability of the trained network. The primary cause for this lack of generalization can be attributed to the nonrepresentative nature of the training data. Table 4.1 shows the mean and variance of the formant data used to train the BP network described in the previous chapter. These formants were extracted from isolated utterances of vowels corresponding to the source and target speakers. In the table the columns marked M and V refer to the mean and variance, respectively. From the table we can observe that the variance of formant frequencies extracted from isolated utterances of vowels is very small. In the following section we show that by using a more representative training set, the generalization capability of the network can be improved.

4.3 Improving generalization – Use of representative data set

This section describes how the generalization capability of a network can be improved by using a formant training data set which represents the required formant transformation operation effectively. Our aim is to train a network in such a way that the trained network must be able to transform formant transitions besides transforming steady formants. A straightforward method is to train the network with formants extracted from nonsteady regions of speech also, for example, formants extracted from vowel to vowel ti-ansition regions. A major problem in using formants extracted from nonsteady regions of speech is that of finding correspondences between the formants extracted from the speech of the source speaker and those from the target speaker. This problem can be circumvented using dynamic time warping(DTW) algorithm [33] to compute the correspondences between frames. But experiments showed that the DTW algorithm can give wrong correspondences which may affect adversely the training of the network [23]. The following section describes the use of a training set which will improve the generalization capability of the network and also circumvent the problem of determining correspondence.

4.3.1 Speech data for improving generalization

Continuous sentences uttered by both the source and the target speakers were segmented manually to mark steady vowel regions. The first three formants extracted from the frame having the maximum energy in each of the steady vowel regions constitutes the data for the present study. In this way formant data was collected from fifty sentences. We had nearly five hundred pairs of formant vectors for training the neural network. The advantage of using such a data set is that the natural variability of the formants are captured in the data set. Table 4.2 shows the mean and variance of this clata set. Comparing Table 4.1 and Table 4.2 we note that the variance of

Table 4.2: Mean and variance of the formant frequencies extracted from steady vowels occurring in continuous speech.

Vowels	Source speaker						Target speaker					
	F_1		F_2		F_3		F_1		F_2		F_3	
	M	V	' M	V	M	V	M	V	M	V	M	V
/ī/	687	86	1334	92	2379	145	873	92	1556	115	2773	73
/ē/	475	54	1980	89	2674	72	590	114	2218	200	2890	122
/ā/	324	26	2194	59	2792	59	403	28	2586	94	3173	132
/ō/	534	81	1226	104	2459	96	628	126	1232	135	2735	158
/ū/	372	31	1191	116	2436	85	462	46	1176	139	2894	75

the training data,set has significantly increased when formants were extracted from vowels occurring in sentences. For example this increase is significant in the case of F_3 corresponding to the vowel /ī/ (compare seventh columns of Table **4.1** and Table **4.2)** and F_2 of /ē/ (compare eleventh column of Table **4.1** Table 4.2). Thus we expect this training data set to help the network in improving the generalization capacity.

4.3.2 Training procedure

A multilayer feedforward neural network with two hidden layers was used to capture the implicit nonlinear formant transformation function. The network consists of three elements each in the input and output layers, and eight elements in each of the hidden layers. The formant data used for this study is same as that described in Section **4.3.1.** Figure **4.1** shows the way in which this network was trained to capture the required transformation. In the training phase we present formant values extracted from the steady vowel regions of the source speaker. The desired

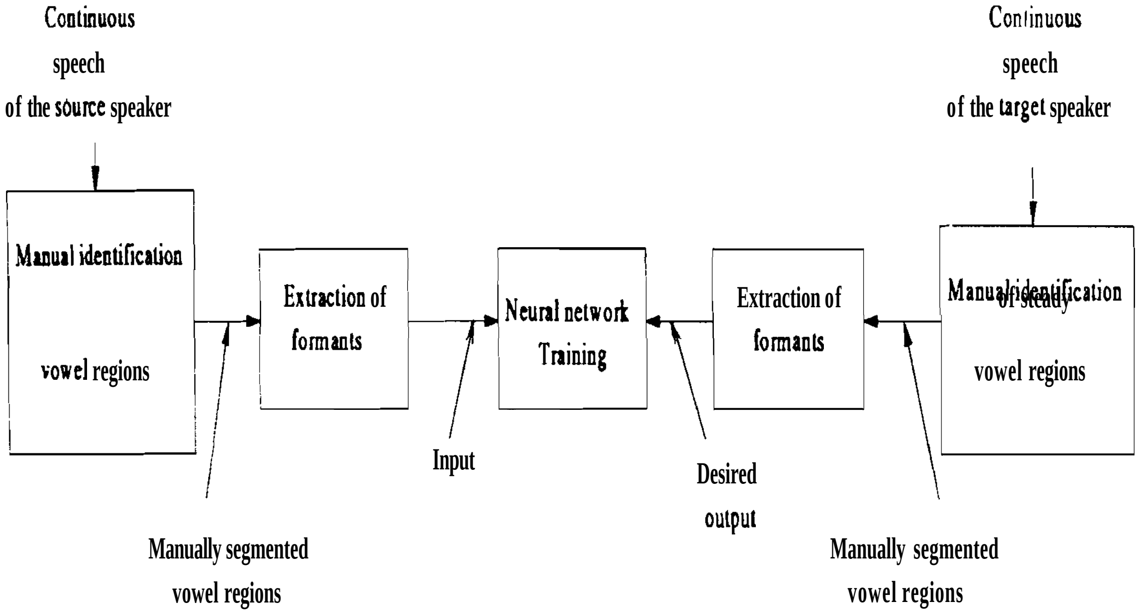


Figure 4.1: Training procedure of a feedforward network using the formants extracted from steady vowels occurring in continuous speech.

output is the formants extracted from the corresponding steady vowel region of the speech of the target speaker. After training, it is expected that the neural network would have captured a function which is capable of transforming not only steady vowels but also formant transitions.

For capturing the formant transformation function, we are training the network with formants extracted from steady vowel regions. But we expect the trained network to he able to transform formant transitions as well. That is, we expect the trained network to transform input formant vectors for which it has not been trained. This implies that the network must be trained in such a way that it learns enough from the training set, so that it can generalize what it has learned from the training set. In order to achieve good generalization, it must be ensured that the training does not cause an overfitting of the input-output data set. A statistical tool called cross validation [32] was used to train the network in such a way that it doesn't overfit the training data [32]. In this method training set is partitioned into two sets, (i) a set for estimation and (ii) a set for validation. The network is trained using the data in the estimation set. After each iteration the network is tested on the data in the validation set. The error in the output of the network for input data in the training set is termed as the training error and the error given on the validation set is the generalization error. Figure 4.2 shows the plot of the training and the generalization errors versus iteration number. From the figure it can be observed that even though the training error reduces, the generalization error shows a rising trend after a certain number of iterations. This indicates that after a certain number of iterations the network parameters are adjusted to overfit the data in the training set. Hence the training is stopped when the generalization error starts increasing, even though further training will result in a reduced training error.

4.3.3 Testing generalization capability of the network

In this section we describe a procedure to evaluate the generalization capability of the trained network in the context of transforming formant contours. In the specific problem of formant transformation, generalization capability of the network refers to the ability to transform formant transitions without introducing discontinuities or other distortions, even though the network is trained using formants extracted from steady sounds (vowels). We can extract formants from speech utterances of the source speaker corresponding to vowel sequences and transform it using the trained network. The transformed formant contour can then be compared with the formant contour for the target speaker corresponding to the same sound segments. The error between the transformed and the target formant contour will give a measure of the generalization capacity of the network. But there are two major problems in using formant transitions extracted from natural speech. Firstly, the source and the target speech will be typically warped in time and thus we can't directly compare the target

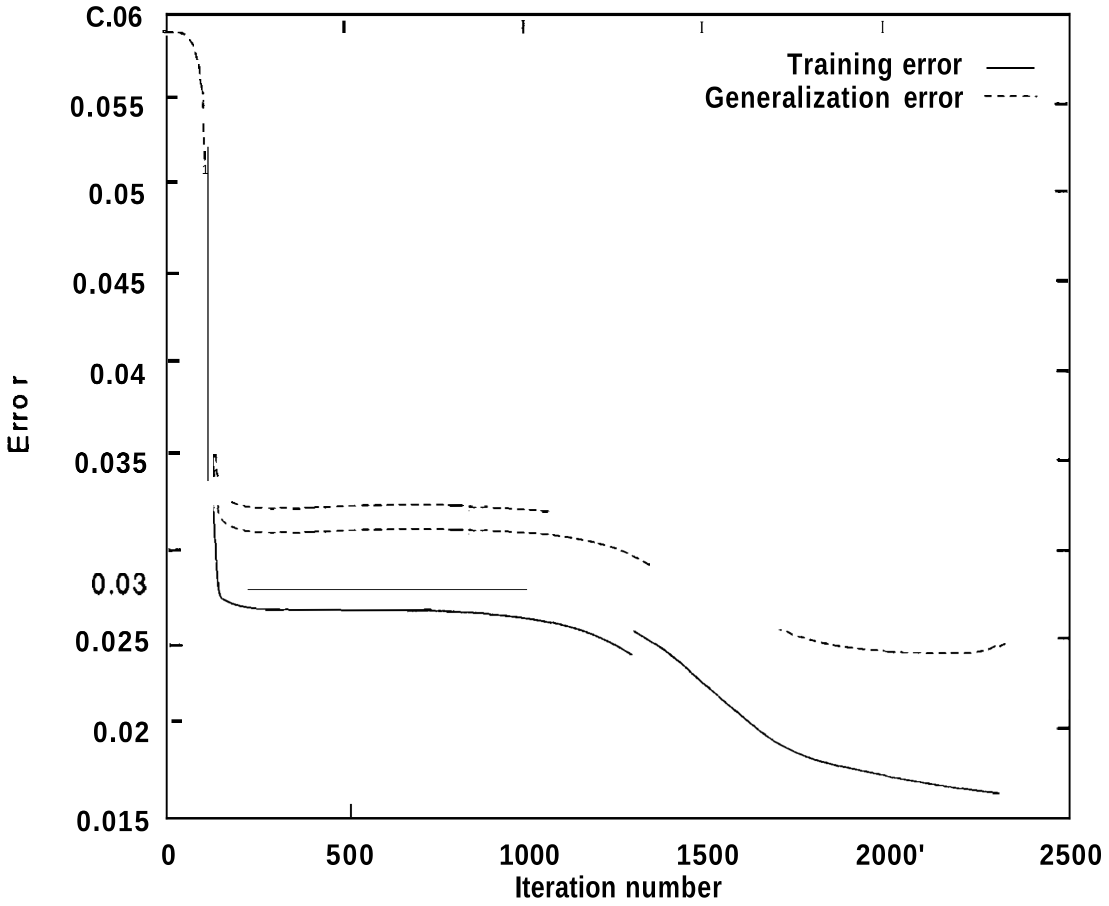


Figure 4.2: Variation of the training and generalization errors vary as training progresses.

and the transformed formants. Moreover, there may be individual variations in the formant trajectories clue to interspeaker variations and coarticulation, which the network may not have captured. Thus, in order to test the generalization capability of the trained network we propose to use synthetic formant contours.

Assume that the source and target training set consists of formant data represented by F^i_{vs} for i = 1, 2 ... N and F^i_{vt} for i = 1, 2 ... N where v can be one of the five vowels /ā/, /ē/, /ī/, /ō/ or /ū/ and N represents the number of formant data set used for training. We compute the mean formant vector corresponding to the various vowels for both the source and the target speakers in the following way.

$$\bar{F}_{vs} = \frac{1}{N}\sum_{i=1}^{N} F^{i}_{vs}$$

$$\bar{F}_{vt} = \frac{1}{N}\sum_{i=1}^{N} F^{i}_{vt}$$

A source formant transition, for example, corresponding to the vowel sequence /āī/, is derived by interpolating $\bar{F}_{as}$ and $\bar{F}_{is}$ by a monotonically increasing/decreasing function. We represent this synthetic formant contour by $F^{\vec{n}}_{ai_s}$, for n = 1,2 . . . l, where 1 is the number of points used to interpolate $\bar{F}_{as}$ and $\bar{F}_{is}$. The corresponding target formant is derived by interpolating $\bar{F}_{at}$ and $\bar{F}_{it}$ by the same function which was used to interpolate the source mean formant vectors, and is represented by $F^{\vec{n}}_{ai_t}$, for n = 1, 2 . . . l. Now the synthetic source formant transition $F^{n}_{ai_s}$ corresponding to the vowel sequence /āī/, is transformed using the trained network. The transformed formant transition is represented by $F^{\vec{n}}_{ai_{tr}}$, for n = 1, 2 . . . l. The generalization error for the formant transition corresponding to the vowel sequence /āī/ is given by

$$GE_{ai} = \frac{1}{l}\sum_{n=1}^{l} \frac{F^{\vec{n}}_{ai_{tr}} - F^{\vec{n}}_{ai_t}}{max(F^{\vec{n}}_{ai_{tr}}, F^{\vec{n}}_{ai_t})}. \qquad (4.1)$$

The total generalization error is given by adding the generalization error corresponding to all possible vowel to vowel formant transitions. Figure **4.3** gives the detailed algorithm of the proposed generalization test. We have conducted the above test of generalization on a network trained using formants extracted from isolated utterances of steacly vowels and also on a network,trained using the formant data describecl in Section 2.3. Table **4.3** shows the generalization errors for the two networks, one trainecl using isolated utterances of steady vowels and the other using the steady vowels extracted from continuous speech. The results show a significant reduction in the generalization error for the second case. Figure **4.4** shows the results obtained in transforming formants extracted from the source speaker corresponding to the vowel sequences /āī/, /āū/ and /ōī/. Hence it is clear that, even though the network was trained with formants extracted from the steady vowels occurring in continuous sentences, it has faithfully transformed formant transitions as well.

```
    The training set consists of
        $F^i_{vs}, F^i_{vt}$        i = 1, 2 ... N
Step I: Calculation of the mean formant vectors corresponding to the five vowels
        for all v
        begin
            $\bar{F}_{vs} = \frac{1}{N} \sum_{i=1}^{N} F^i_{vs}$
            $\bar{F}_{vt} = \frac{1}{N} \sum_{i=1}^{N} F^i_{vt}$
        end
Step II: Computation of all the possible vowel to vowel formant transitions
        for all combinations of v$_1$ and v$_2$
        begin
            $\vec{F}^n_{v_1v_2s} = CS(\bar{F}_{v_1s}, \bar{F}_{v_2s}, n)$        n = 1, 2 $\cdots$ $l$
            $\vec{F}^n_{v_1v_2t} = CS(\bar{F}_{v_1t}, \bar{F}_{v_2t}, n)$        n = 1, 2 $\cdots$ $l$
            CS represents the cubic spline interpolation operation
        end
Step III: Transformation of the source formant transitions
        for all possible combinations of v$_1$ v$_2$
        begin
            $\vec{F}^n_{v_1v_2tr} = f(\vec{F}^n_{v_1v_2s})$        $n = 1, 2 \cdots l$
            *f(.)* is the function learned by the network
        end
Step IV: Calculation of the generalisation error
        for all possible combinations of v$_1$ and v$_2$
        begin
            $GE_{v_1v_2} = \frac{1}{l} \sum_{n=1}^{l} \frac{\vec{F}^n_{v_1v_2tr} - \vec{F}^n_{v_1v_2t}}{max(\vec{F}^n_{v_1v_2tr}, \vec{F}^n_{v_1v_2t})}$
        end
```

Figure 4.3: Algorithm for testing the generalization capability in the context of formant transformation.

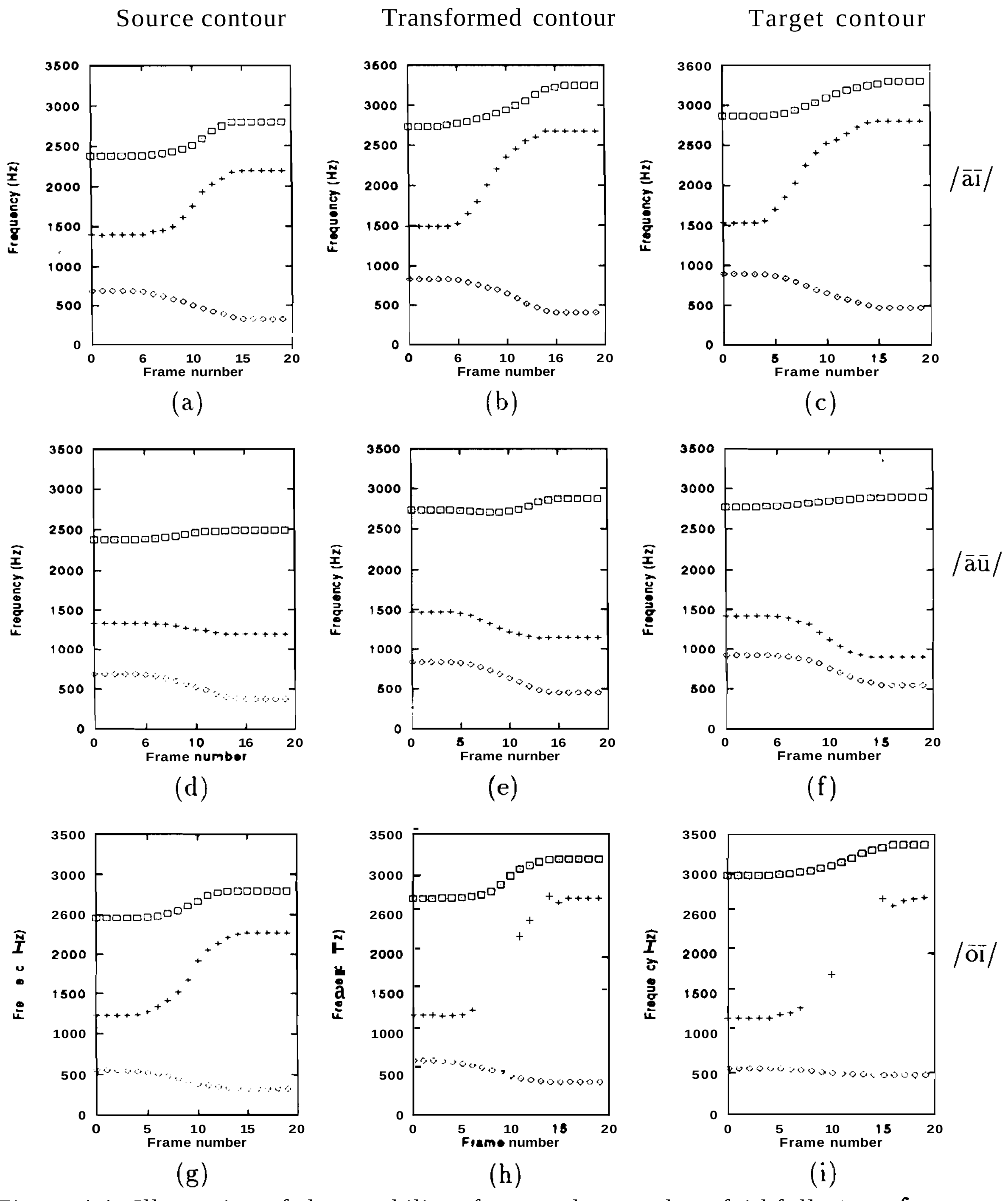


Figure 4.4: Illustration of the capability of a neural network to faithfully transform formant transitions. (a),(d) and (g) show the formant contour corresponding to the vowel sequences /āī/, /āū/ ancl /ōī/ extracted from the speech of the source speaker (male). (b),(e) ancl (h) show the corresponding transformed formant contour using the trained network. (c),(f) and (i) show the corresponding target formant contours (female).

Table 4.3: Table showing the capability of a representative training data set in improving the generalization capability of a network.

Vowel sequences	Generalization error given by a network trained using formants extracted from isolated utterances of vowels			Generalization error given by the network trained using formants extracted from vowels occurring in continuous speech		
	F_1	F_2	F_3	F_1	F_2	F_3
/āē/	8.8	11.8	5.0	1.7	6.6	2.6
/āī/	13.9	11.8	4.4	1.8	5.8	3.1
/āō/	3.3	11.7	7.2	2.1	4.1	1.4
/āū/	6.7	15.5	3.6	1.9	6.1	3.0
/ēī/	8.7	7.3	5.8	1.3	5.0	1.0
/ēō/	6.9	13.3	6.5	1.8	3.1	1.4
/ēū/	11.0	11.8	4.0	0.6	2.4	1.4
/īō/	7.6	14.0	5.6	1.9	3.8	1.9
/īū/	11.2	11.4	3.1	1.0	4.3	1.3
/ōū/	26.8	23.1	7.9	1.2	3.0	1.5

4.4 Summary

In Chapter 3 we had shown that a neural network trained using formants extracted from steady vowels is not capable of transforming formant transitions faithfully . The reason for the failure is due to lack of generalization capability of the network. In this chapter we have demonstrated that by using representative training data set the generalization capability of the network can be improved significantly. In this context we have proposed a method to measure the generalization capacity of a trained network. From the results of the experiments in this chapter it can be concluded that a BP network trained using formants extracted from steady vowels occurring in continuous speech is capable of transforming not only the steady formants but formant transitions also. In the next chapter we describe a method to incorporate the formant transformation learned by a neural network into a voice' conversion system.

Chapter 5

Implementation of voice transformation

5.1 Introduction

There are two phases in the development of a voice conversion system, a learning phase and a transformation phase. In the learning phase various factors that are responsible for voice personality are identified and the speaker specific knowledge is acquired and represented in a proper form. In the transformation phase the given speech signal is modified using the knowledge acquired during the learning phase. In this chapter we describes the various issues in the development of the transformation phase. We focus particularly on the incorporation and testing of the vocal tract system transformation. We first use the transformed formants directly to synthesize speech with a formant synthesizer. Then the segmental quality of the transformed speech will be poor due to lack of bandwidth information and also due to errors in the extraction of formants. We propose a method in which we use a trained neural network to modify the LPCs extracted from the speech of the source speaker. These modified LPCs are then used to synthesize the transformed speech. In ortler to perceive the quality of voice conversion we have to incorporate source characteristics also into the speech. We have used a simple linear transformation to modify the average pitch. But interspeaker variations are not limited to the segmental level. Hence for evaluating the quality of the transformation done at the segmental level we have to mask the speaker characteristics at the suprasegmental level. We propose an algorithm to normalize the intonational features between two speakers. Finally, we test the performance of the proposed voice transformations between several pairs of speakers.

The following section tlescribes the voice conversion system. Section **5.3** proposes an algorithm for modifying the LPCs extracted from the speech of the source speaker to effect the vocal tract system transformation. Section 5.4 emphasizes the need for normalizing interspeaker variations in the suprasegmental features for evaluating the quality of the voice transformation derived using information at the segmental level. Section 5.4.1 briefly mentions the features of intonation patterns with reference to Hintli. The algorithm to normalize these patterns between speakers is described in Section **5.4.2.** Experiments for evaluating the quality of voice conversion are described in Section 5.5.

5.2 Voice transformation system

The transformation phase of voice conversion involves: (i) Extraction of speaker dependent parameters from the speech of the source speaker and (ii) modifying these parameters to match those of the target speaker. The modification is done using the speaker-dependent knowledge acquired during the learning phase. After the modification of the speaker-dependent parameters, speech in the voice of the target speaker is synthesizetl using the modified parameters.

Figure 5.1 gives the block diagram of the transformation phase of voice conversion. The first operation done on the speech signal is a voiced/unvoiced labeling. After this preliminary analysis, parameters are extracted from the source speech sig-

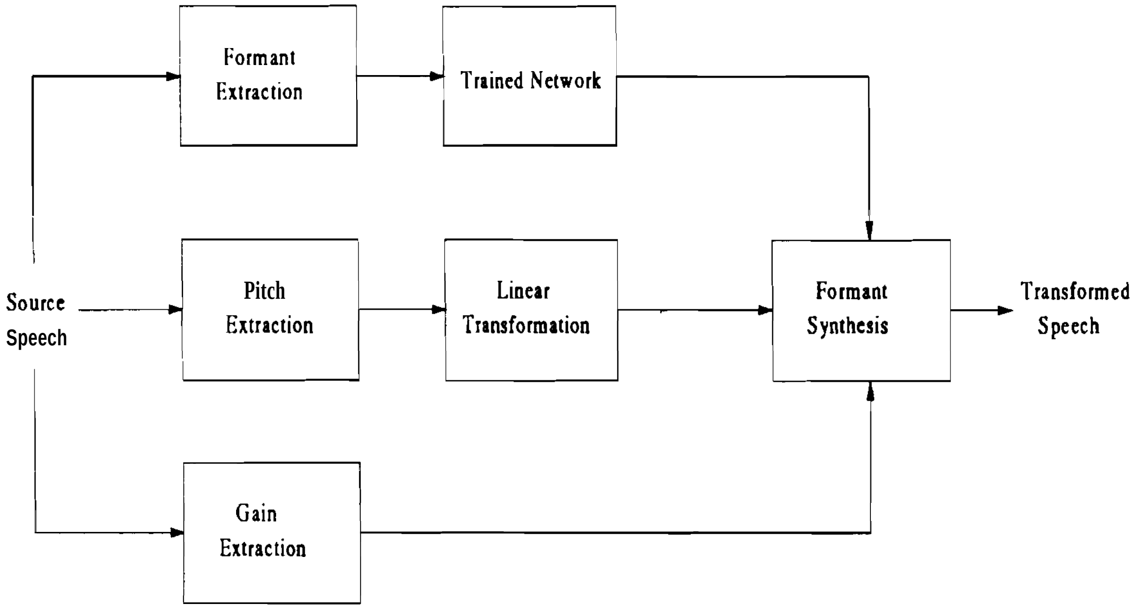


Figure 5.1: The block diagram showing the transformation phase of voice conversion.

nal. For parameter extraction we have used a sliding window of duration 25.6 msecs and a shift of 6.4 msecs. The parameters are pitch, energy and formants. The SIFT algorithm [29] was used to extract pitch. The first three formants are extracted using an algorithm based on the properties of minimum phase group delay functions [27]. The extracted parameters are transformed to incorporate the characteristics of the target speaker. It is reasonable to assume that speaker specific information is mainly in the voicecl segments of speech [9]. Hence only those parameters which are extractetl from voicetl frames are modified.

5.3 Incorporation of formant transformation

As shown in the Figure 5.1, the formant transformation can be incorporatetl in a straight forward manner. A direct transformation of formants and the use of the transformed formants to synthesize speech will lead to poor quality in the synthe-

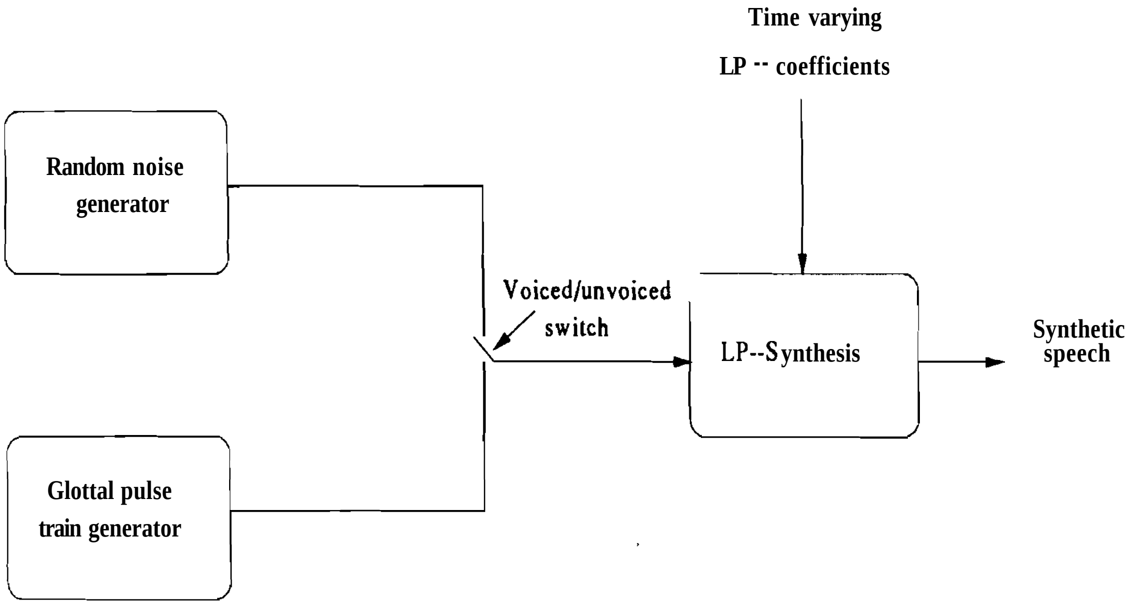


Figure 5.2: The basic structure of an LPC – vocoder.

sised speech. The reasons for the poor quality of synthetic speech is due to lack of bandwidth information and unreliability in the estimated formants. In this context, we propose to use an LPC – vocoder [11], since it produces better quality speech. We discuss the salient features of an LPC - vocoder. Then we describe an algorithm to modify the LPCs extracted from the speech of the source speaker. Figure 5.2 shows the basic structure of an LPC – vocoder. The vocal tract system is represented by a time-varying digital filter. This filter is specified by the LP coefficients as follows:

$$H(z) = \frac{1}{A(z)} = \frac{G}{1 + \sum_{i=1}^{p} a_i z^{-i}}, \tag{5.1}$$

where p is tlie order of the all-pole system. This filter is excited with random noise during unvoiced frames and with a train of periodic glottal pulses during voiced frames to generate synthetic speech. The roots of the linear prediction polynomial $A(z)$ will have real and complex conjugate roots. The complex conjugate roots represent the vocal tract systenr resonance (formants). Suppose a complex conjugate root is represented hy ref^{ℓ}, then the corresponding vocal tract resonant frequencies

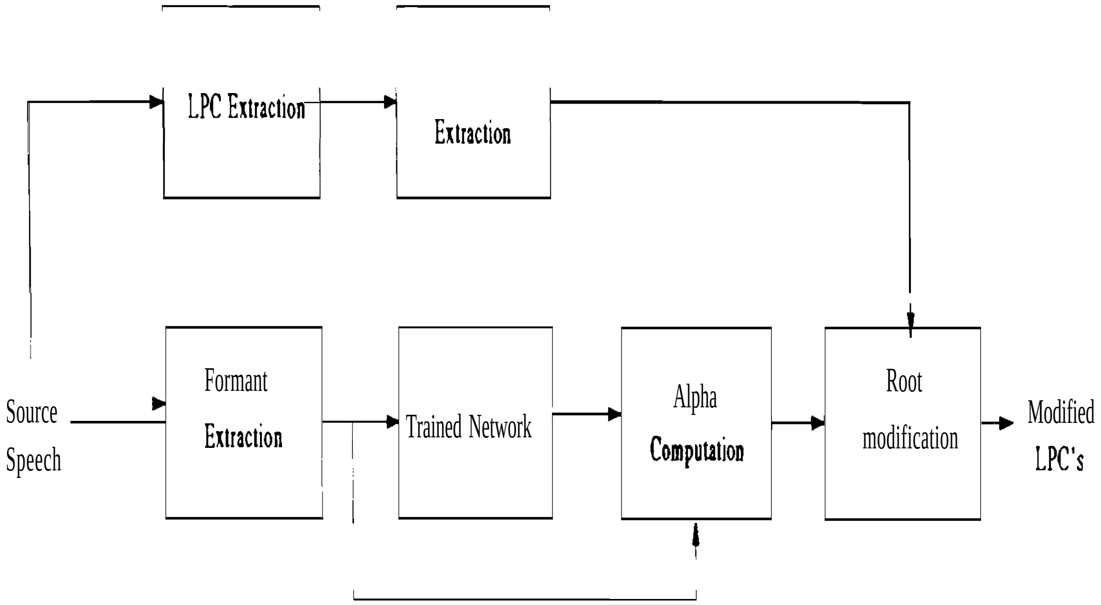


Figure 5.3: Block diagram showing the various steps involved in the modification of LPCs using a trained network.

and bandwidths are given by

$$
\begin{aligned}
f &= \frac{\theta}{2\pi T}, & (5.2)\\
b &= \frac{-\log r}{\pi T}, & (5.3)
\end{aligned}
$$

where T is the sampling period. From equation (5.3) it is evident that θ is directly proportional to the formant frequency. Hence by shifting the complex pole in the z-plane we can implement a formant transformation. Figure 5.3 shows the block diagram of the LPC modification algorithm. LPCs (10^{th} order) are extracted from the speech of the source speaker along with the other parameters described in Section 5.2. Suppose the source formants extracted from a frame of speech are represented by F_s^i, for i = 1, 2 3. These source formants are transformed using the trained neural network to get the transformed formants represented by F_t^i, for i = 1, 2, 3. From the source and the transformed formants we compute a set of formant scale factors given by α^i, for $i = 1, 2, 3$, where

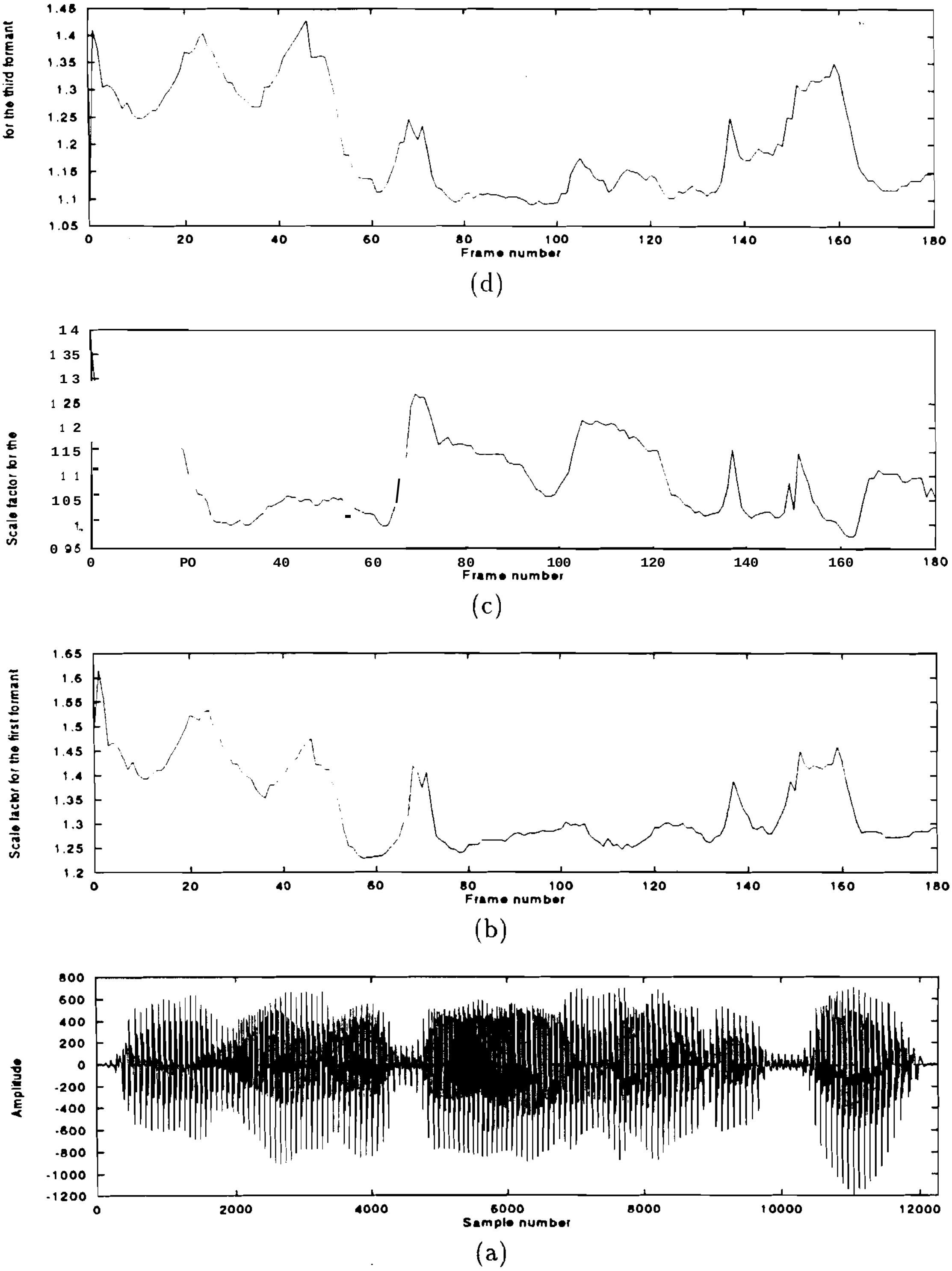
$$\alpha^i = \frac{F^i_t}{F^i_s} \qquad (5.4)$$

α^i represents the scaling factor corresponding to the i^th^ formant. Thus for each frame we get a set of αs which gives the amount by which the source formants extracted from that frame need to be scaled. It must be noted that the *as* vary significantly across a sentence. This variation is shown in the Figure **5.4.** The figure shows the α contour corresponding to the all voiced sentence "We were away a year ago". This also shows the highly nonlinear nature of the formant transformation between speakers. These scale factors are used to modify the LPCs extracted from each of the frames.

First the roots of the linear prediction polynomial are obtained. The complex conjugate pole pairs correspond to the vocal tract resonances. In order to effect a vocal tract system transformation these complex roots are modified or shifted using the scale factors. In this procedure the real roots are left unaltered. Suppose that $r^i e^{\theta^i}$ is a complex root corresponding to the i^th^ formant, where r corresponds to the formant bandwidth and **0** corresponds to the formant location (center frequency). Thus using the formant scale factors we will be modifying only the **Os,** using the following equation

$$\begin{aligned} \theta^i_{mod} &= \left[1 + \left(1 - \alpha^i\right)\frac{\pi - \theta^i}{\pi}\right]\theta^i \qquad if\ \alpha^i > 1 && (5.5) \\ \theta^i_{mod} &= \alpha^i\theta^i \qquad if\ \alpha^i \le 1 && (5.6) \end{aligned}$$

If there are more than three complex conjugate pairs of roots, then the complex roots corresponding to the fourth and higher formants are modified in a similar way by using the scale factor corresponding to the third formant. The detailed algorithm used for the modification of the LP-roots is given in Figure 5.5. The way in which the complex roots are shifted in the z-plane is illustrated in the Figure 5.6.



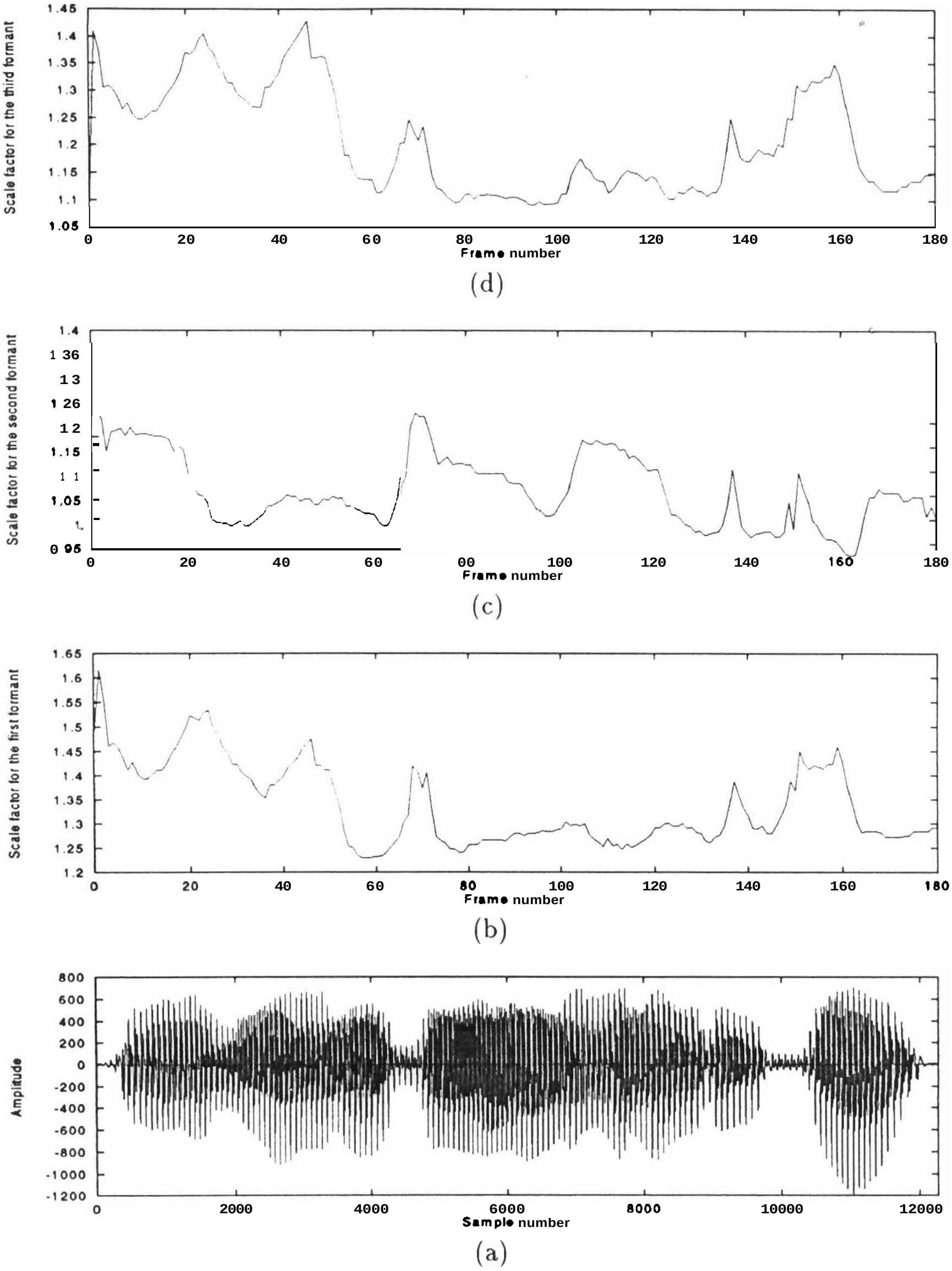
(d)

(c)

(b)

(a)

Figure 5.4: Illustration of alpha contour. (a) Speech waveform corresponding to the text "We were away a year ago". (b),(c) and (d) show the variation of α_1,α_2 and α_3 across the sentence.



(d)

(c)

(b)

(a)

Figure 5.4: Illustration of alpha contour. (a) Speech waveform corresponding to the test "We were away a year ago". (b), (c) and (d) show the variation of α_1,α_2 and α_3 across the sentence.

The source formant vectors are given by

$F_s^i \qquad i = 1, 2 \ldots 3$

Step I

Transformation of the source formant vectors using the trained network:

$\bar{F}_{tr} = f(\bar{F}_s)$

where $f()$ represents the function learned by the network

Step II

Computation of the scale factors

$\alpha^i = \frac{F_{tr}^i}{F_s^i}$

Step III

Root solving of the LPC polynomial

$r^i e^{j\theta^i}$ represents a complex pole corresponding to the i^{th} formant

Step IV

Root modification

$r_{mod}^i = r^i$

$\theta_{mod}^i = \left[1 + (1 - a')\frac{\pi - \theta^i}{\pi}\right]\theta^i \qquad \text{if } \alpha^i > 1$

$\theta_{mod}^i = \alpha^i \theta^i \qquad \text{if } \alpha^i \leq 1$

Step V

Recomputation of the LPCs from the modified roots

Figure 5.5: Algorithm to modify the LPCs using the trained network.

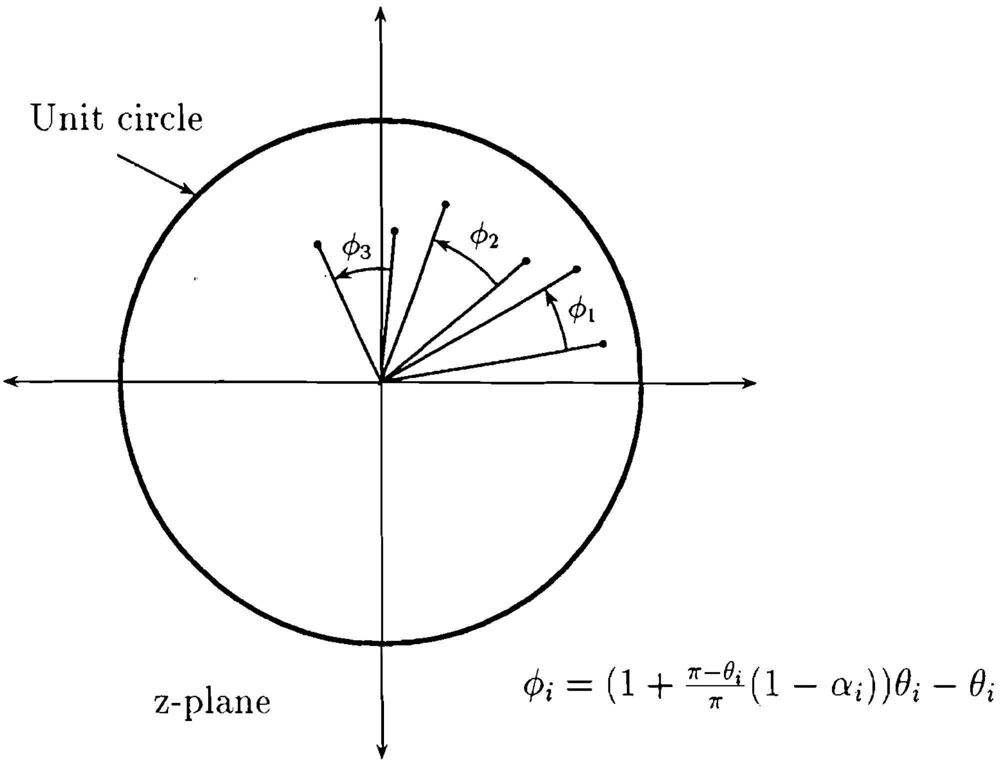

Unit circle
ϕ_3
ϕ_2
ϕ_1
z-plane
$\phi_i = (1 + \frac{\pi-\theta_i}{\pi}(1 - \alpha_i))\theta_i - \theta_i$

Figure 5.6: Illustration of the way in which the poles are shifted in the z-plane.

5.4 Normalization of intonational features

In orcler to assess the quality of the voice transformation we compare the voice quality of transformed sentences with the same sentences uttered by the target speaker. Since the converted speech is synthetic we cannot compare it with the natural utterance of the target speaker. Hence we synthesize speech from the pitch, gain ancl LPCs extracted from the test utterance of the target speaker using an LPC vocoder. By informal listening we compare the transformed speech with the speech synthesized using the parameters extracted from the target speaker's utterance. In such a comparison, the interspeaker variations at the suprasegmental level must be eliminated from the test utterance, in order to make a good judgement on the effectiveness of the transformation done at the segmental level. At the suprasegmental level intonation plays a major role in providing an individuality to the speech of a speaker. Hence we propose a method by which the intonational characteristics in a sentence uttered by both the source and the target speakers are normalized automatically. The aim is to modify the intonation pattern of the test utterance of the target speaker so that it matches with that of the source speaker. In order to normalize the intonation pattern of the target speaker we use the general features of the intonation patterns of Hindi sentences [34, 35], which are briefly described in the following section. The discussion is applicable only for utterances in Hindi.

5.4.1 Characteristics of intonation patterns in Hindi

This section presents a model for intonation patterns in Hindi [34]. The important components in the description of intonation patterns in Hindi are: (i) Declination, (ii) Local fall-rise, (iii) Resetting and tapering effect [34, 36]. These features independently and collectively represent important linguistic information and characterize an individual's voice. Figure 5.7 illustrated the intonation pattern of a typical declarative sentence:

ātmā amar hai *šarīr nāsvān* hai
soul immortal is body mortal is — (literal translation)
The soul is immortal and the body is mortal — (meaning)

The F_0 which sets off (about 115 Hz) from the onset of the periodicity of the signal assumes the maximum F_0 level (about 170 Hz) at the final syllable of the first word (*/-mā/* in*/ātmā/*). The F_0 contour drifts down from this point towards another on the initial syllable of the next content word (*/a- /* in */amar/*) to about 115 Hz. Again, it rises towards a higher point (about 130 Hz) in the final syllable (*/-mar/* in */amar/*) of the word. The F_0 contour falls off towards a lower point (about 105 Hz) and rises towards another point within the same word (*/hai/*) and finally it tapers

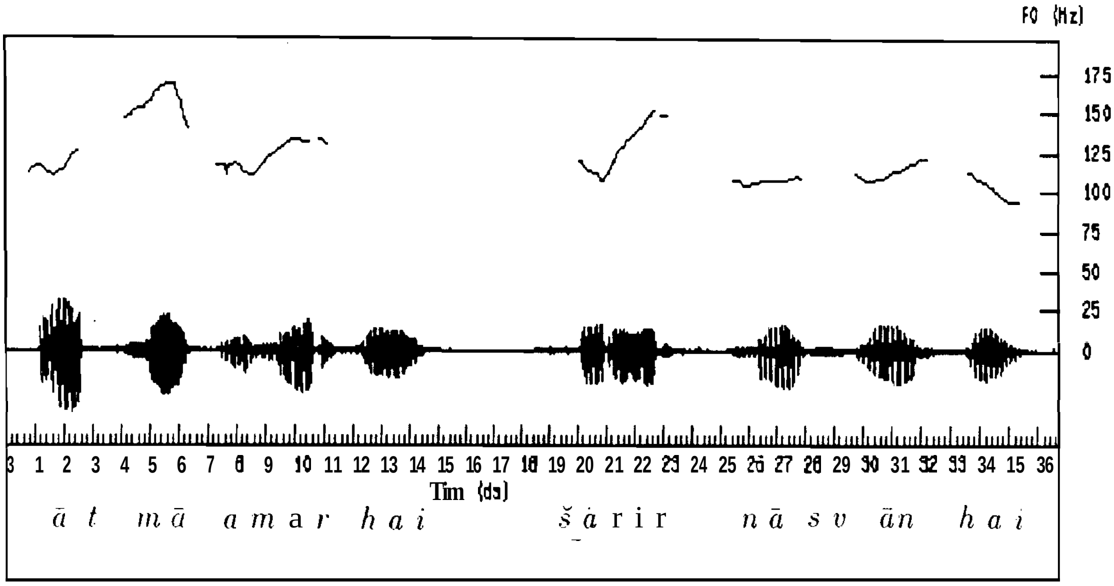


Figure 5.7: Pitch pattern of a typical declarative sentence.

off to about 110 Hz. The F_0 contour ol the utterance is characterized by a few target *points.* The target points are the local maxima and minima of F_0 which result in rise and fall of F_0 movements. The local minima and maxima are called valleys and peaks respectively. They are connected by transition lines. If two imaginary grid lines are drawn in a declarative sentence, one connecting all the peaks and the other all the valleys, it is possible to say that the F_0 contour drifts down as a function of time till the occurrence of a major syntactic or semantic break (at the end of*/ātmā amar hai/* and at tlie end of tlie sentence), which is also marked by a significant pause of a duration of about 300 ms. The grid lines show an upward trend in the case of interrogative sentences. The valleys and peaks alternate each other till the entl of the sentence. They may occur within the region of a syllable or across syllables. This is called fall-rise pattern and this is determined by the phonological patterns of tlie constituent words of tlie utterance and other linguistic factors, The difference between the F_0 values measuretl at a valley and the following peak is called the F_0

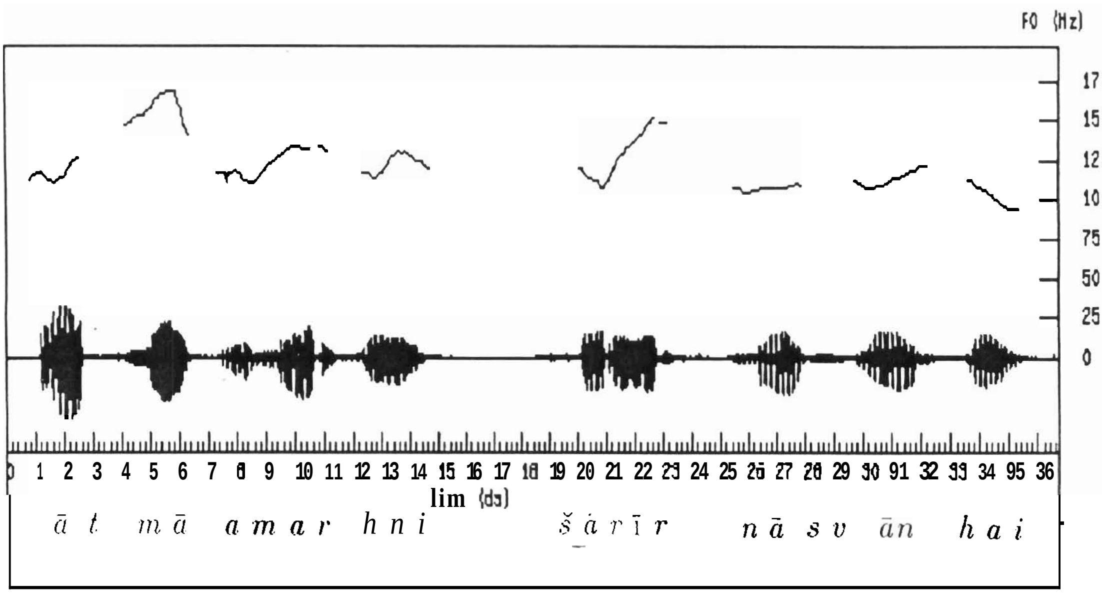


Figure 5.7: Pitch pattern of a typical declarative sentence.

off to about 110 Hz. The F_0 contour of the utterance is characterized by a few *target* points. The target points are the local maxima and minima of F_0 which result in rise and fall of F_0 movements. The local minima and maxima are called *valleys* and *peaks* respectively. They are connected by *transition lines.* If two imaginary grid lines are drawn in a cleclarative sentence, one connecting all the peaks and the other all the valleys, it is possible to say that the F_0 contour drifts down as a function of time till the occurrence of a major syntactic or semantic break (at the end of*/ātmā amar hai/* and at the end of the sentence), which is also marked by a significant pause of a duration of about 300 ms. The grid lines show an upward trend in the case of interrogative sentences. The valleys and peaks alternate each other till tile encl of the sentence. They may occur within the region of a syllable or across syllables. This is called fall-rise pattern and this is determined by the phonological patterns of the constituent words of the utterance and other linguistic factors. The difference between the F_0 values measured at a valley and the following peak is called the F_0

range. The F_0 range is another important feature which carries a lot of speaker specific information and speaking style [37, **38,** 39].

If we assume that the rate of fall of F_0 values is constant, then we can model the valleys and peaks as points on two separate lines, the base line and the top line, respectively. Thus if P_1 and P_2 are the two peak F_0 values measured at times T_1 and T_2, then the equation of the top line [40] becomes:

$$P(t) = P_1 + \frac{P_2 - P_1}{T_2 - T_1}(t - T_1). \tag{5.7}$$

Thus, once we model the pitch contour using the above equation, it is possible to predict the F_0 value of any peak if we know the position(time) at which it occurs. Similarly the base line is modeled by the following equation.

$$V(t) = V_1 + \frac{V_2 - V_1}{V_2 - V_1}(t - T_1). \tag{5.8}$$

where V_1 and V_2 are the F_0 values of two valleys, and T_1 and T_2 are the time instants at which these valleys occur.

5.4.2 An algorithm for intonation normalization

Figure **5.8** shows the block diagram of the proposed intonation normalization algorithm.

From the pitch contours of the target and the source utterances the vowel nuclei are identified using an algorithm described in [35]. The pitch values at this vowel nuclei is considered as the saddle points of the entire pitch contour. These F_0 values corresponding to the source and the target formants is represented by F_{0s}^i for $i = 1, 2 \ldots N$ and F_{0t}^i for $i = 1, 2 \ldots N$, respectively. N is the number of syllables in the sentence. Note that perceptually significant feature in an intonation pattern is the relative values of the F_0 measured at the vowel nuclei and not the absolute F_0 values. Hence we modify the range that is defined by F_0 values measured at the vowel nucleus of successive syllables.

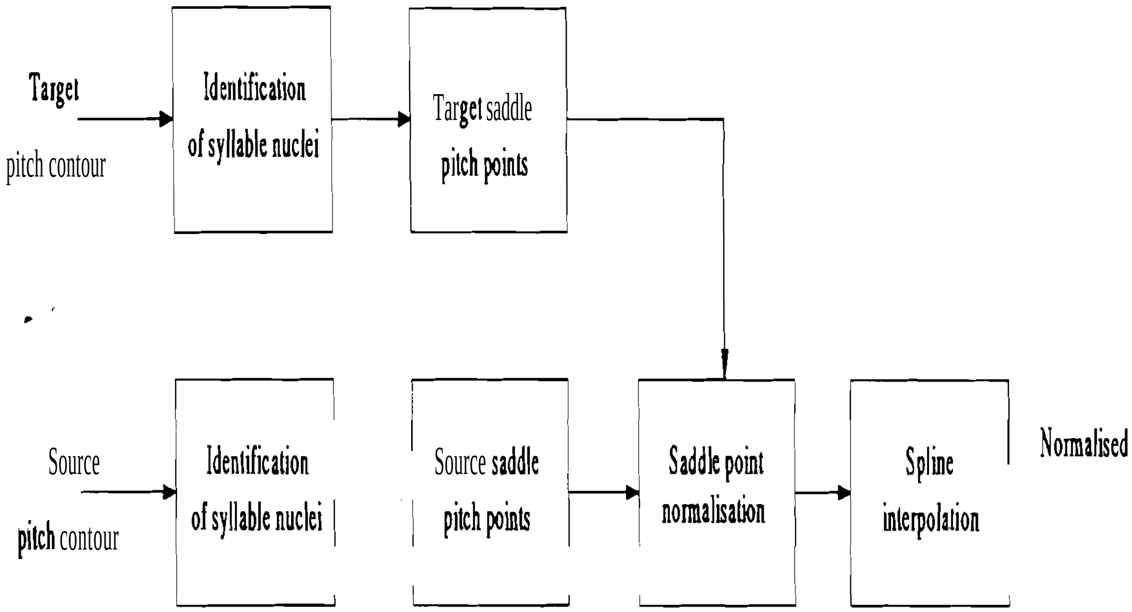


Figure 5.8: Block diagram showing the various steps involved in the proposed intonation normalization algorithm.

F_{0t}^{1} is not modified. F_{0t}^{2} is modified so that the range between F_{0t}^{1} and F_{0t}^{2} measured in semitones becomes equal to the range between F_{0s}^{1} and F_{0s}^{2}. This modification is continued till we modify the range between F_{0t}^{N-1} and F_{0t}^{N}. From the modified saddle points, F_0 values of the other voiced frames are computed using a cubic spline interpolation [41]. The detailed algorithm is given in the Figure 5.9

5.5 Evaluation of voice transformation

In this section we describe the voice conversion studies performed between pairs of speakers. The following voice conversions were carried out.

Case A: Male to Female

Case B: Female to Male

Case C: Male to Male

Case D: Female to Female

```
Step I
    Extraction of the source and the target $F_0$ contours
Step II
    Identification of the saddle points using the algorithm described in [35]
      The $F_0$ saddle points are given by
        $F^i_{0s}$ .    $i = 1, 2 \ldots N$
        $F^i_{0t}$      $i = 1, 2 \ldots N$
        Where N is the number of syllables in the sentence
Step 111
    Modification of $F^i_{0t}$
      $F^1_{0t_{norm}} = F^1_{0t}$
      fori = 2 ... N
        begin
            $F^i_{0t_{norm}} = F^{i-1}_{0t} \frac{F^i_{0s}}{F^{i-1}_{0s}}$
        end
Step IV
    Construction of the entire pitch contour
      Construct the normalized pitch contour from the modified saddle points
      by cubic spline interpolation
```

Figure 5.9: Algorithm for intonation normalization scheme.

The basic objective is to asses the quality of the transformation for all the above cases. Speech data corresponding to fifty sentences spoken by two male and two female speakers were collected. Formant transformation corresponding to the above described types of transformations were captured as mentioned in Chapter **4.** The Corresponding linear pitch transformations were also obtained. The learned formant and pitch transformations were used in the four cases of voice conversion. The transformed speech was obtained for the following conditions.

1. ***Average pitch modification:*** Speech with modified average pitch and original vocal tract system characteristics of the source speaker.

2. ***Formant transformation:*** Speech with original pitch and the transformed formants.

3. ***Average pitch and formant transformation:*** Speech with modified average pitch and transformed formants.

These were compared with the speech synthesized from the natural utterances of the target speaker. The pitch contour used to synthesize this was normalized to match with that of the source pitch contour using the algorithm described in Section **3.** Informal listening shows that for the third case which includes both formant transformation and the average pitch modification does indeed bring in the characteristics of the target speaker in the synthesized speech.

It was noted that whenever the target speaker is a female, the conversion quality becomes poor. This is consistent with the observations in [14, 42]. The reason for this can be attributed to the general problems in synthesizing female speech [43]. The quality of voice conversion was found to degrade in the order shown in the Figure 5.10.

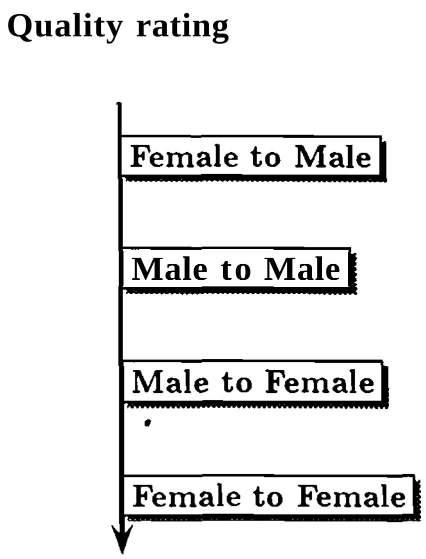


Figure 5.10: Figure showing the way in which the quality of voice conversion decreases depending on the type of conversion being attempted.

5.6 . Conclusion

In this chapter we have described a system for incorporation of vocal tract transformation to realize a voice conversion. We have observed that by combining formant transformation and LP-synthesis we can develop an effective way of transforming the vocal tract system characteristics of the source speaker to match with those of the target speaker. This method uses both the formant and LPC representations of the vocal tract system characteristics. The need for normalizing the suprasegmental characteristics for evaluating the effectiveness of conversion has been discussed. We have also proposed an algorithm to normalize the intonation patterns of the transformed and the target speech. Informal listening showed that the quality of transformation is highest while converting female voice to male voice. We have observed a degradation in the quality of the transformed speech when the target speaker is female.

Chapter 6

Conclusion

6.1 Summary

In this thesis we have addressed issues related to the problem of voice conversion. The various factors responsible for voice characteristics were discussed. Since the speaker characteristics at the linguistic and suprasegmental levels are learned features, it is difficult to model interspeaker variations at these levels and capture them as transformations. Voice characteristics at the suprasegmental level can be captured only by manual analysis of large amount of speech data. However speaker characteristics at the segmental level can be attributed mainly to variations in the characteristics of the vocal tract system and thus can be modeled as a transformation operations. The specific objective of the work was to capture the transformation of the vocal tract system characteristics between two speakers, so that the speech of the source speaker can be transformed or modified to incorporate the features of the target speaker.

We have used linear function approximation to capture the formant transformation corresponding to steady vowels. We have found that even for steady vowel sounds the formant transformation is highly nonlinear. Approximating the transformation by piecewise linear transforms results in discontinuities in the formant contours. Moreover, to apply this piecewise linear transformation to any set of formants, we should know the vowel class from which the formants were extracted.

A multilayer feedforward neural network with nonlinear processing elements can capture any arbitrary input output mapping function. We have proposed such a network to capture the implicit formant transformation function across the source and the target formants. We have demonstrated that such a network is capable of reducing some of the problems in linearly approximating the formant transformation. Even though a neural network is capable of transforming formant transitions without introducing discontinuities into the formant contour, the transformed formant transitions were steeper than the target formant transitions. The failure of the neural network for capturing the formant transitions shows the lack of generalization capability of the trained network. The main reason for this lack of generalization is due to poor representation of the transformation information in the training data set. The generalization capability of a network can be improved by using a more representative set of training data. For this we have used formants extracted from the steady vowel regions occurring in continuous speech for training the network. The advantage of using such a training data set is the variability introduced into the data set which will help improve generalization. We have demonstrated that a feedforward network trained with the above mentioned data set provides an improvement in generalization capability and thus can transform formant transitions faithfully. In this context we have suggested a method for measuring the generalization capability of a network trained to capture the nonlinear formant transformation, by testing the network using synthetic formant transitions.

Finally we have discussed issues in incorporating the formant transformation into a voice conversion system. Since it is easier to synthesize better quality speech with LPCs, we have proposed a method for modifying the LPCs to transform source formants to match with those of the target. This method of transforming the vocal tract system characteristics takes into account the advantages of two different methods of representing the vocal tract system, namely, using LPC and using formants.

6.2 Future directions

The objective of the thesis was to capture the implicit nonlinear transformation of the vocal tract system characteristics across speakers. We have focused only on the transformation of formant frequencies, where as formant bandwidths also contribute to the voice characteristics [44]. Thus extraction and transformation of formant bandwidths can improve the quality of the transformation. Similarly nasalization is another factor which is used by humans in differentiating speakers from their voices. The dynamics of the glottal source also contributes to voice characteristics. If one could extract parameters corresponding to these features and transform them reliably, then the quality of voice transformation will improve significantly. The main difficulty in the use of these features for voice conversion is lack of reliable algorithms to extract parameters corresponding to these features.

Even though in this work we have focused only on the segmental aspects of voice conversion, the intricate speaker characteristics lie at higher levels of knowledge (linguistic and suprasegmental levels). Moreover humans extensively use the speaker characteristic5 at these levels for identifying speakers from their voices. Hence for accomplishing the task of voice conversion it is very important to analyze and model the speaker characteristics at the linguistic and suprasegmental levels.

We have estimated the quality of the voice transformation by informal listening tests. A challenging problem in voice conversion is to develop an objective measure of voice characteristics.

As far as voice conversion at the segmental level is concerned, it is evident that the limit on the quality of conversion is set by lack of reliable methods for extracting speaker dependent parameters from speech on the one hand and by the quality of the synthetic speech on the other hand. Thus it can be concluded that any improvement in the field of parameter extraction or speech synthesis will lead to an improvement in the quality of voice transformation.

Bibliography

[1] J. Lyons, *Semantics.* Cambridge: Cambridge University Press, 1977.

[2] F. Nolan, *The Phonetic Bases of Speaker Recognition.* Cambridge: Cambridge University Press, 1983.

[3] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliff, NJ: Prentice Hall, 1993.

[4] D. O'Shaughnessy, *Speech Communication – Human and Machine.* Massachusetts: Addison-Wesley, 1987.

[5] G. R. Doddington, "Speaker recognition-identifying people by their voices," *Proc.* IEEE, vol. 11, pp. 1651–1664, 1985.

[6] B. S. Atal, "Automatic recognition of speakers from their voices," *Proc.* IEEE, vol. 6, pp. 460–475, 1976.

[7] H. Gish and M. Schmidt, "Text–independent speaker identification," IEEE *ASSP Magazine,* vol. 11, pp. 460–475, 1994.

[8] N. Iwahashi and Y. Sagisaka, "Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks," *Speech Communication,* vol. 16, pp. 139–151, 1995.

[9] J. P. Eatock and J. Mason, "A quantitative assessment of the relative speaker discrimination properties of phonemes," in Proc. IEEE Internat. Conf. Acoust. Speech Signal Processing, pp. 1133–1136, 1994.

[10] J. P. Eatock and J. Mason, "Automatically focusing on good discriminating speech segments in speaker recognition," in Proc. *Internat.* Conf. on Spoken Language Processing, pp. 133–136, 1990.

[11] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," IEEE Trans. Acoust. Speech Signal Processing, vol. 50, pp. 637–655, 1971.

[12] S. Seneff, "System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction," IEEE Trans. Acoust. Speech Signal Processing, vol. 30, pp. 566–578, 1982.

[13] D. G. Childers, B. Yegnanarayana, and K. Wu, "Voice conversion: Factors responsible for voice quality," in Proc. IEEE Internat. Conf. Acoust. Speech Signal Processing, pp. 19.10.1–19.10.4, 1985.

[14] D. G. Childers, K. Wu, D. M. Hicks, and B. Yegnanarayana, "Voice conversion," Speech Communication, vol. 8, pp. 147–158, 1989.

[15] G. Fant, "Glottal flow: Models and interaction," *J.* of Phonetics, vol. 14, pp. 393–399, 1986.

[16] J. Slifka and T. R. Anderson, "Speaker modification with LPC pole analysis," in Proc. IEEE Internat. Conf. Acoust. Speech Signal Processing, pp. 644–647, 199.5.

[17] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in Proc. IEEE Internat. Conf. Acoust. Speech Signal Processing, pp. 655–658, 1988.

[18] M. Savic and I. H. Nam, “Voice personality transformation," Digital Signal Processing, vol. 1, pp. 107–110, 1991.

[19] H. Mizuno and M. Abe, "Voice conversion based on piecewise linear conversion rules of formant frequency and spectrum tilt," in Proceedings of the *1994* International Symposium on Speech, Image Processing and Neural Networks, (Hong Kong), pp. 1469–1472, 1994.

[20] H. Mizuno and M. Abe, "Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt," Speech Communication, vol. 16, pp. 153–164, 1989.

[21] M. Abe, K. Shikano, and H. Kuwabara, “Cross language voice conversion," in Proc. IEEE Internat. Conf. *Acoust.* Speech Signal Processing, pp. 345–348, 1990.

[22] M. Abe, "A segment based voice conversion," in Proc. IEEE Internat. Conf. Acoust. Speech Signal Processing, pp. 765–768, 1991.

[23] H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA technique," Speech Communication, vol. 11, pp. 175–187, 1992.

[24] E. Moulines and 3. Laroche, "Non–parametric techniques for pitch-scale and time-scale modification of speech," Speech Communication, vol. 16, pp. 175–205, 1995.

[25] E. Moulines and F. Charpentier, “Pitch–synchronous waveform processing techniques for text–to–speech synthesis using diphones," Speech Communication, vol. 9, pp. 453–467, 1990.

[26] B. Widrow and S. Stearns, Adaptive Signal Processing. Englewood Cliff, NJ: Prentice–Hall. 1985.

[27] H. A. Murthy and B. Yegnanarayana, "Formant extraction from minimum phase group delay function," Speech *Communication*, vol. 10, pp. 209–221, 1991.

[28] G. Fant, A. Kruckenburg, and L. Nord, "Prosodic and segmental speaker variations," Speech Conzmunication, vol. 10, pp. 521–531, 1991.

[29] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans.* Audio Electroacoust., vol. 20, pp. 367–378, 1972.

[30] K. Hornik, M. Stinchcombe, and H. White, "Multilayer networks are universal approximators," Neural Networks, vol. 2, pp. 359–366, 1989.

[31] T. L. McClelland, D. E. Rumelhart, and the PDP Research Group, Parallel Distributed Processing. Cambridge, MA: MIT Press, 1986.

[32] S. Haykin, Neural Networks. New York: Macmillan, 1994.

[33] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE* Trans. Acoust. Speech Signal Processing, vol. 26, pp. 43–49, 1978.

[34] A. S. Madhukumar, S. Rajendran, and B. Yegnanarayana, "Intonation component of a text to speech system for Hindi," Computer Speech and *Languagε*, vol. 7, pp. 283–301, 1993.

[35] S. Rajendran and B. Yegnanarayana, "Word boundary hypothesization for continuous speech in Hindi based on F_0 patterns," To appear in *Speech Communi*cation.

[36] D. R. Ladd, "Intonational phrasing: The case for recursive prosodic structure," Phonology Yearbook, vol. 3, pp. 311–340, 1986.

[37] B. L. Brown, W. J. Strong, and A. C. Rencher, "Fifty-four voices from two: the effect of simultaneous manipulations of rate, mean fundamental frequency and variance of fundamental frequency on rating of personality from speech," J. Acoust. *Soc.* Amer., vol. 55, pp. 313–318, 1974.

[38] K. Scherer, "Personality markers in speech," in Social Markers in Speech (K. R. Scherer and H. Giles, eds.), pp. 147–209, Cambridge: Cambridge University Press, 1979.

[39] C. G. Henton, "Fact and fiction in the description of female and male pitch:" Language *&* Communication, vol. 9, pp. 299–311, 1989.

[40] A. S. Madhukumar, Intonation Knowledge for Speech Systems *for* an Indian Language. PhD. Thesis: Indian Institute of Technology, Madras, 1993.

[41] D. Rogers and J. A. Adams, Mathematical elements of computer graphics. Ney York: McGraw Hill, 1989.

[42] N. B. Pinto, D. G. Childers, and A. L. Lalwani, "Formant speech synthesis: Improving production quality," *IEEE* Trans. Acoust. Speech *Signal* Processing, vol. 37, pp. 1870–1887, 1989.

[43] I. Karlsson, "Female voices in speech synthesis," Speech Communication, vol. 19, pp. 111–120, 1991.

[44] D. G. Childers and K. Wu, "Gender recognition from speech: Part II:fine analysis," *J.* Acoust. *Soc.* Amer., vol. 90, pp. 1841–1856, 1991.

List of Figures

List of Tables

List of publications and reports

Technical papers

1. M. Narendranath, Hema A. Murthy, S. Rajendran and B. Yegnanarayana, "Voice conversion using artificial neural networks," Proc. *Workshop on* Automatic Speaker Recognition, Identification and *Verification*, Switzerland, April, 1994.

2. M. Narendranath, Hema A. Murthy, S. Rajendran and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," Speech Communication, vol.16, no.2, pp.206–216, 1995.

Technical reports

1. B. Yegnanarayana, M. Narendranath and S. Rajendran, *Voice Conversion*, Department of Computer Science and Engineering, Indian Institute of Technology, Madras, September 1995 (Technical report submitted to the Department of Science and Technology, Govt. of India).

2. B. Yegnanarayana, Hema A. Murthy, S. Rajendsan and M. Narendranath, "Voice conversion - An Overview", Technical Report No. 1, Department of Computer Science and Engineering, Indian Institute of Technology, Madras, August 1993.

3. B. Yegnanarayana, Hema A. Murthy, S. Rajendran and M. Narendranath, "Voice conversion - Different approaches”, Technical Report No. 2, Department of Computer Science and Engineering,Indian Institute of Technology, Madras, August 1993.