ELSEVIER

Speech Communication 16 (1995) 207–216

SPEECH COMMUNICATION

Transformation of formants for voice conversion using artificial neural networks

M. Narendranath, Hema A. Murthy, S. Rajendran, B. Yegnanarayana *

Department of Computer Science and Engineering, Indian Institute of Technology, Madras 600 036, India

Received 24 May 1994; revised 22 November 1994

Abstract

In this paper we propose a scheme for developing a voice conversion system that converts the speech signal uttered by a source speaker to a speech signal having the voice characteristics of the target speaker. In particular, we address the issue of transformation of the vocal tract system features from one speaker to another. Formants are used to represent the vocal tract system features and a formant vocoder is used for synthesis. The scheme consists of a formant analysis phase, followed by a learning phase in which the implicit formant transformation is captured by a neural network. The transformed formants together with the pitch contour modified to suit the average pitch of the target speaker are used to synthesize speech with the desired vocal tract system characteristics.

Zusammenfassung

In diesem Artikel wird ein Konzept für die Entwicklung eines Stimmumwandlungsystems vorgestellt, daß das Sprachsignal eines Ausgangssprechers in das den Eigenschaften des Zielsprechers entsprechende Sprachsignal umwandelt. Dabei werden im Besonderen die Umwandlungsergebnisse der Merkmale des Vokaltraktsystems von einem Sprecher zum anderen transferiert. Für die Darstellung des Vokaltraktsystems werden die Formanten benutzt, für die Sprachsynthese ein Formantvocoder. Das Konzept besteht aus einer Formantanalysephase, gefolgt von einer Lernphase, in der die implizite Formantumwandlung von einem Neuronennetzwerk aufgegriffen wird. Die umgewandelten Formanten, zusammen mit den modifizierten, der durchschnittlichen Grundfrequenz des Zielsprechers angepaßten Grundfrequenzverläufen werden für die Sprachsynthese benutzt, die die angestrebten Merkmale des Vokaltraktsystems aufweist.

Résumé

Dans cet article, nous décrivons une méthode de transformation du timbre de la voix. Ce dispositif permet de transformer certains paramètres de la voix d'un locuteur source de façon à approcher les caractéristiques acoustiques de la voix d'un locuteur cible. Nous nous intéressons ici particulièrement à la transformation du spectre à court-terme. Dans cette contribution, nous utilisons comme paramètre les formants, connus pour représenter de façon satisfaisante les caractéristiques acoustiques du conduit vocal. Notre méthode de transformation comprend

* Corresponding author. E-mail: yegna@iitm.ernet.in.

SSDI 0167-6393(94)00058-1

deux phases distinctes: une phase d'analyse, dans laquelle nous extrayons les paramètres formantiques, et une phase d'apprentissage dans laquelle nous apprenons les transformations à l'aide d'un réseau de neurones. Les formants transformés sont ensuite utilisés, lors de la synthèse, dans un synthétiseur à formants.

Keywords: Voice conversion; Speaker characteristics; Formants; Multilayer feedforward neural network

1. Introduction

Speech signal possesses mainly two kinds of information, namely the speech message part and the speaker identity part. Isolating the characteristics of speech and speaker from the signal is a challenging problem in speech research. Extracting the message part of the information is the focus of research in the area of speech recognition (Rabiner and Juang, 1993). Speaker recognition and verification deals with techniques to extract speaker dependent information from the speech signal.

In the development of a voice conversion system mainly two problems are to be addressed. They are (1) identification of speaker characteristics or acquisition of speaker dependent knowledge in the analysis phase and (2) incorporation of the speaker specific knowledge while synthesis during the transformation phase. This is relevant in two situations. For example, in a text-to-speech system it may be required to generate speech with the desired voice characteristics. Voice conversion is also relevant while transforming speech from a source speaker into a speech signal with voice characteristics of the target speaker. Both these involve identification of speaker characteristics, and extraction of these characteristics from the speech signal. These characteristics need to be represented in a suitable manner for incorporating them either in a text-to-speech system or in a voice transformation system.

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 necessary for speaker normalization for developing speaker independent speech recognition systems. Several attempts have been made to study and manipulate the speaker dependent characteristics. Atal and Hanauer (1971) studied the feasibility of modifying voice characteristics using an LPC vocoder. Seneff (1982) demonstrated a method to modify the excitation and vocal tract parameters. Childers et al. (1985, 1987, 1989) have examined methods for converting the speech of a male speaker to sound like that of a female speaker and vice versa. Abe et al. (1988) have developed a technique for voice conversion through vector quantization and spectral mapping. In a similar work reported by Savic and Nam (1991) the mapping code book was realized by a neural network. Another work reported by Abe (1991) describes a voice conversion algorithm that uses speech segments as conversion units. To produce speech in a different voice the corresponding segments are replaced. In another method for voice conversion technique proposed by Valbret et al. (1992), prosodic modifications were incorporated in the excitation signal using PSOLA (Pitch Synchronous Overlap Add) technique and speech was synthesized using the transformed spectral parameters.

In this paper we train a neural network to learn a transformation function which can transform the speaker dependent parameters extracted from the speech of the source speaker to match with that of the target speaker. In particular, we address the issue of transforming the characteristics of the vocal tract system in the speech signal of the source speaker to that of the target speaker. In the voice conversion system that is discussed in this paper, we assume a formant vocoder model for speech production. In Section 2 we discuss a general scheme for voice conversion, and the issue of transformation of the vocal tract characteristics. In Section 3 we propose a neural network model for capturing the implicit transformation of the formants representing the vocal tract systems of two speakers. A synthesis experiment incorporating the formant transformation is described in Section 4 to demonstrate the applicability of the neural network model for voice transformations.

2. Speaker characteristics for voice conversion

In this section we first identify parameters which characterize inter-speaker variations and then develop methods for transforming them across speakers.

The term speaker characteristics or voice is used to refer to those factors in the spoken utterance which carry information about the speaker, i.e., those factors which are used by listeners to identify the speaker of an utterance. In general these factors are not known precisely. But still listeners are able to distinguish among speakers from their voices without much effort. Studies in interspeaker variations and factors affecting voice quality have revealed that there are various parameters in the speech signal, both at the segmental and at the suprasegmental level, which contribute to the interspeaker variability (Klatt and Klatt, 1990; Fant et al., 1991; Childers and Lee, 1991).

From a speaker recognition point of view, one would also use this knowledge at various levels to recognize a speaker from his voice. At the highest level voices are differentiated by the use of linguistic cues derived from the speech. These linguistic cues include the language of the speaker, his dialect, choice of lexical patterns, choice of syntactic constructs and the semantic context. The characteristics of a speaker at the linguistic level are difficult to analyze and model.

There are factors in a spoken utterance which are speaker dependent and can be measured (or estimated) from the acoustic waveform of speech. These factors are the acoustic level characterization of the speaker. The acoustic level characterization can be further divided into segmental and suprasegmental levels. At the segment level the vocal tract system and source characteristics of the speaker contributes to the speaker characteristics. The vocal tract may be characterized by a linear time varying system represented by a set of time varying parameters. From parameter extraction point of view, it is convenient to represent the system as a linear digital filter, for example an all-pole model. However, from a transformation point of view, it is convenient to represent the system with articulatory parameters. Since articulatory parameters are difficult to extract from the speech signal, as a compromise, formants are proposed for representing the vocal tract system information. The nature of the source, especially for voiced segment, is an important feature of speaker characteristics. Besides the pitch period, the shape of the glottal pulse is also unique for a given speaker. In this study we consider transformation of the average pitch only. We assume a standard model for the glottal pulse for excitation instead of deriving the transformation of the glottal pulse characteristics, although such a transformation of source information is critical for incorporating the characteristics of the voice of the desired speaker in the synthesis.

There are many factors at the suprasegmental level which affect the speaker's voice quality. The prosodic features such as pitch, duration and intensity as a function of time are unique for a given speaker. In fact a speaker is identified more by these factors than by the vocal tract features, as is evident while listening to the linear prediction residual signal of speech. But these features are dependent on the context and also on the language of the speaker. Hence they are more difficult to extract and represent for voice transformation. In the present study suprasegmental features of the source speaker are retained, while using the transformed vocal tract parameters for synthesis.

3. Voice transformation studies

As mentioned before, we focus our attention on the transformation of formants and average pitch of the target speaker in voice conversion. First we study how the formants and average pitch of two speakers differ. We considered five pairs of speakers, each pair consisting of a male and a female speaker. We collected speech data for isolated utterances of vowels /i/, /e/, /a/,

/o/ and /u/ from each of these five pairs of speakers. The first three formants are extracted using a method based on minimum phase group delay functions (Murthy and Yegnanarayana, 1991). Assuming the male speaker as the source (s) speaker and the female speaker as the target (t) speaker, the average values of the formant ratios ($Ki = Fi^t/Fi^s$ for $i = 1,2,3$) are computed for each vowel and for each pair of speakers. The average is obtained across several frames of data from several repetitions of each vowel.

Fig. 1 shows the variation of these scale factors for different vowels, for five different pairs of male and female speakers. These plots show that the scale factors are dependent both on the formant (first, second or third) and the type of the vowel. Moreover, the plots of the three scale factors (corresponding to the three formants) with respect to the various prototype vowels show a similar trend across different sets of male and female speakers.

A notable deviation from the uniform scaling

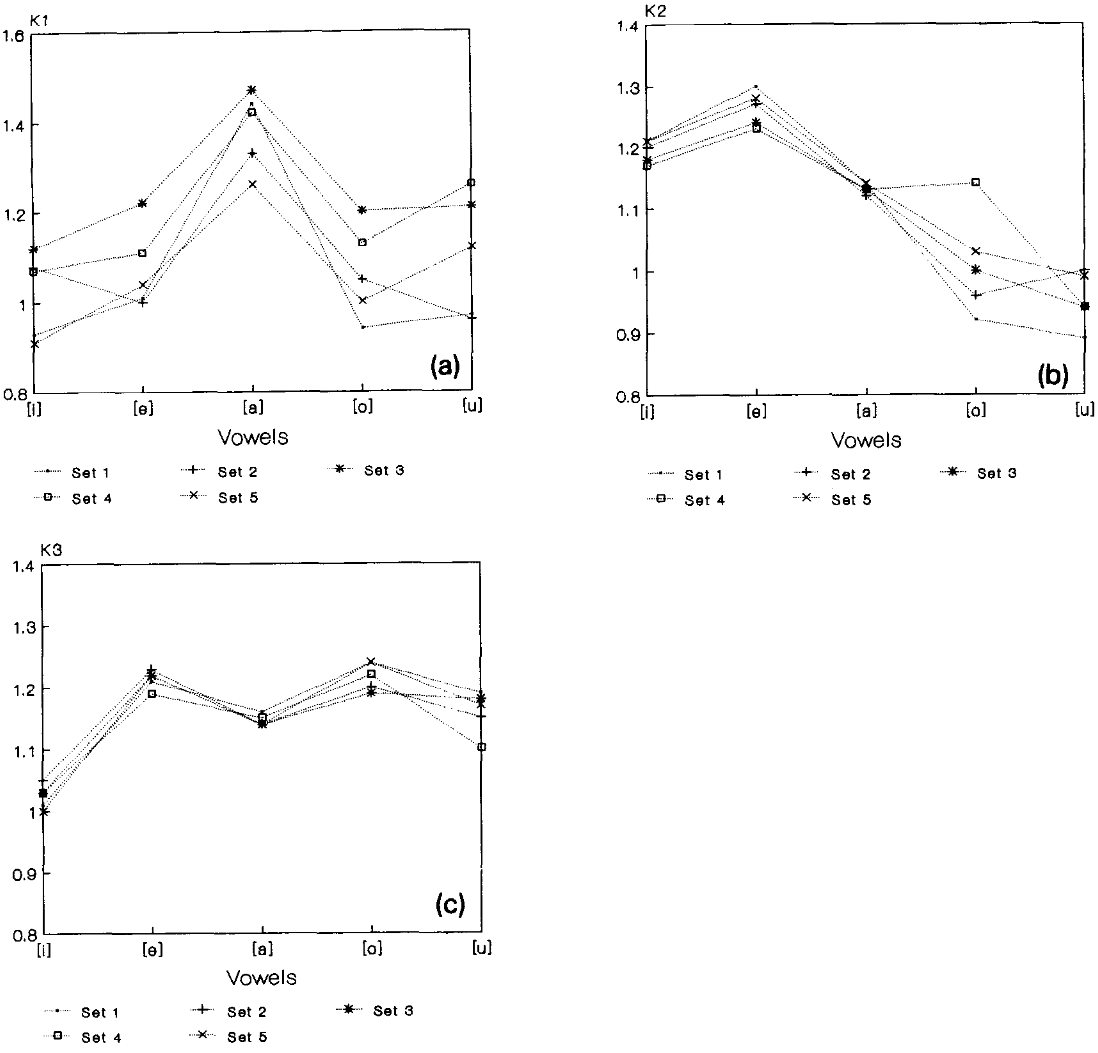


Fig. 1. Formant scale factors for different prototype vowels. These scale factors correspond to a male to female formant scaling. (a) Scale factor $K1$ corrresponding to the first formant. (b) Scale factor $K2$ corrresponding to the second formant. (c) Scale factor $K3$ corrresponding to the third formant.

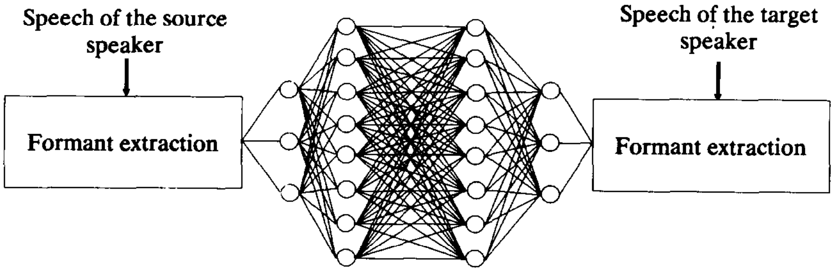


Fig. 2. Training a neural network for capturing a function which transforms the formants of the source speaker to that of the target speaker.

of the formants is the large scale factor for the first formant corresponding to the open vowel /a/ in comparison with the close vowels /u/ and /i/. In the case of the scale factor for the second formant, it is high for front vowels /i/ and /e/. It is worth noting that the back vowels /u/ and /o/ have the second formant scale factor $K2$ that is less than unity. This means that the second formant frequency for back vowels /u/ and /o/ is higher for male speakers than for female speakers. These observations are consistent with a similar study conducted by Fant (Fant et al., 1991).

These plots of the formant ratios show that the variation of the formants from male to female is consistent. That is, one can infer that there is a systematic variation of the vocal tract shapes for all the five pairs of speakers. But this variation is not the same for different vocal tract shapes, as can be seen by the different values of the scale factors for different vowels. This shows that the vocal tract shape transformation between two speakers is not linear.

In order to capture the implicit nonlinear transformation between the vocal tract shapes of two speakers, we propose an artificial neural network model for formant transformation. During the training phase the network is trained with a discrete set of points on the mapping function. If the data set used for training are chosen appropriately then the network will learn a continuous mapping function. Moreover this function can faithfully transform input parameters for which it has not been trained. In continuous speech the vocal tract system characteristics change rapidly across segments. Hence if the transformation involves codebook mapping (Abe et al., 1988; Savic and Nam, 1991), then, for a faithful transforma-

```
repeat
For each set of formant data
 begin
    Step-1:  The formant values (F1-F3) corresponding to the source speaker
             (male) are given as the input.
    Step-2:  The desired output is the formants extracted from the corresponding
             frame of speech of the target speaker (female).
    Step-3:  The weights are adjusted using the backpropagation algorithm.
 end
until the weights converge
```

Fig. 3. Algorithm for training the BP network to capture formant transformation function.

tion, the size of the codebook must be very large. The proposed neural network based transformation system will work for all cases of formant occurrences although, during training, only a few sample pairs of formant vectors are used. This is based on the property that a multilayered feedforward neural network using nonlinear processing elements can capture any arbitrary input–output mapping (Hornik et al., 1989). This generalization property of the neural network helps in the faithful transformation of formants across speakers, avoiding the use of large codebooks.

A network with one input layer (3 units), two hidden layers (8 units each) and an output layer (3 units) is used in this study. The network is trained using the back propagation algorithm to capture the transformation between the formants (McClelland et al., 1986). The data for this study is collected from the steady voiced regions of continuous speech of the source speaker. The corresponding regions from the speech of the target speaker uttering the same sentence are identified manually. The first three formants from these two corresponding steady voiced regions are used as a pair of input and output formant vectors to a neural network. We have collected data from steady voiced regions of speech for a male–female speaker pair for utterances of fifty sentences. We have nearly five hundred such pairs of formant vectors for training a neural network. Fig. 2 shows the block diagram of the neural network system used to capture the transformation function which maps the formants of the source speaker to that of the target speaker. Fig. 3 shows the algorithm used to train the neural network to capture the required transformation.

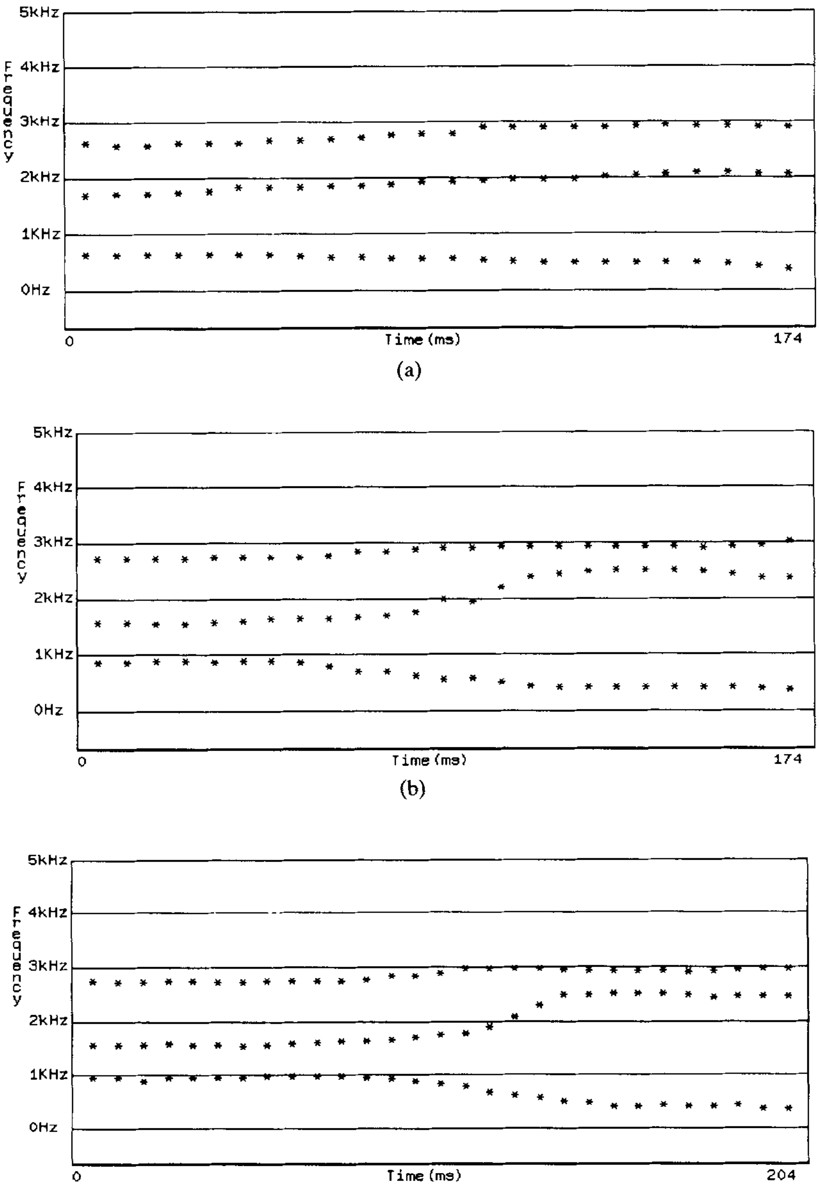
In an attempt to study how far the network has been successful in capturing the relation between the formants of the source and the target speakers, the following two measurements were made:

(1) Table 1 shows the percentage error between the formants for five vowels of the target speaker and the source speaker before and after the application of the transformation learned by the neural network. From the table it is clear that the application of the transformation learned by the neural network to the source speaker's formants brings them closer to those of the target.

(2) The network was trained with formants extracted from the steady voiced regions of continuous speech. But in continuous speech, since the vocal tract changes its shape continuously, the extracted formants will have many transitions. It is therefore desirable that the network be able to transform the formant transitions as well. Fig. 4(a) shows the input formant contours given to the network. The first three formants were extracted from each frame (25.6 ms) of speech data. Fig. 4(b) shows the output of the network or the estimated formant contour. Fig. 4(c) gives the expected target formant contour. Comparing Figs. 4(b) and 4(c) we notice that the formant transition is also transformed well by the network. Therefore we can conclude that although the network is trained only with formants extracted from steady vowel regions, it performs equally well in the regions with for-

Table 1
The percentage error between the source and the target formants before and after the application of the transformation learned by the neural network

Vowels	Error in percentage between the target and the source speaker's formants			Error in percentage between the target and the transformed formants		
	$F1$	$F2$	$F3$	$F1$	$F2$	$F3$
i	15.1	12.3	9.8	5.0	6.2	3.8
e	11.0	15.9	7.8	5.8	5.2	2.8
a	22.0	12.0	13.1	7.3	9.0	5.9
o	12.3	7.9	10.4	7.9	6.0	3.8
u	15.5	10.2	19.3	5.3	6.2	4.6



(a)

(b)

(c)

Fig. 4. The formant contours corresponding to the vowel sequence /a:e:/. (a) Formants extracted from the speech of the source speaker (male). (b) The transformed formants (male to female). (c) Formants extracted from the speech of the target speaker (female).

mant transitions. This illustrates the generalization property of the network which is exploited in the proposed method. This generalization property was verified by comparing the transformed formant contour with the desired formant contours of the target speaker for several sentences.

The pitch frequency for each segment is computed using the SIFT algorithm (Markel, 1972). Fig. 5 shows the variation of the inherent F_0 with

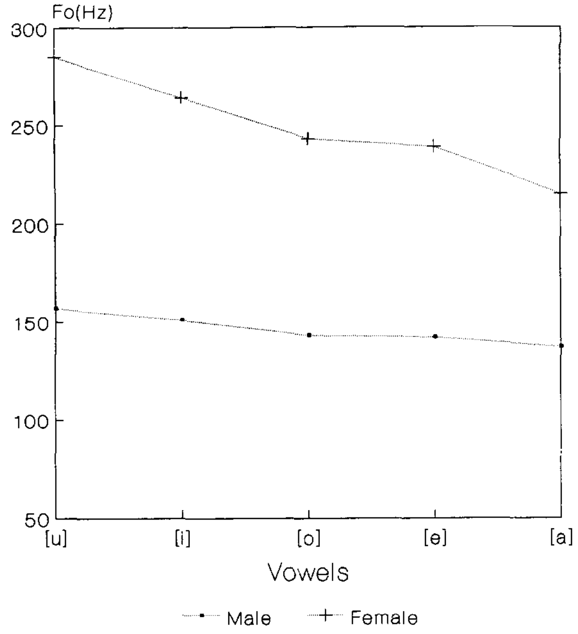


Fig. 5. The variation of F_0 with respect to vowels for male and female speakers.

respect to the five prototype vowels for male and female speakers considered in Fig. 1. The F_0 values were averaged over five male and five female speakers. We can observe a direct relationship between the height of the vowel and the inherent F_0 for both male and female speakers. F_0 is highest for the vowel /u/ and lowest for the vowel /a/. This is true for both male and female speakers. The two main differences in the inherent F_0 of vowels between male and female speakers are:

(a) The average F_0 of female speakers is about 1.6 times higher than that for the male speakers.
(b) The range (max F_0–min F_0) over which the pitch varies for a female speaker is significantly (about three times) larger in comparison with the range for male speakers.

Pitch is extracted from the speech data of the source and target speakers for several voiced segments. The average of the ratio of the pitch frequency of the corresponding steady regions is computed and is used as the pitch modification factor in the synthesis.

4. Synthesis from transformed parameters

Fig. 6 shows the tasks involved in the synthesis phase. Formant transformation is quite straightforward if we have a neural network which has learned the transformation. The three formants extracted from each of the frames of the source speaker's speech are given as input to the input layer of the trained neural network. The output of the network gives the transformed formants.

For modifying the pitch contour of the source speaker we use the ratio of the average F_0 of the source and target speaker. After extracting the speaker dependent parameters from the speech of the source speaker and transforming them, the final step would be synthesizing speech. Formant synthesis was employed for synthesis using the transformed formants and the pitch contour. The

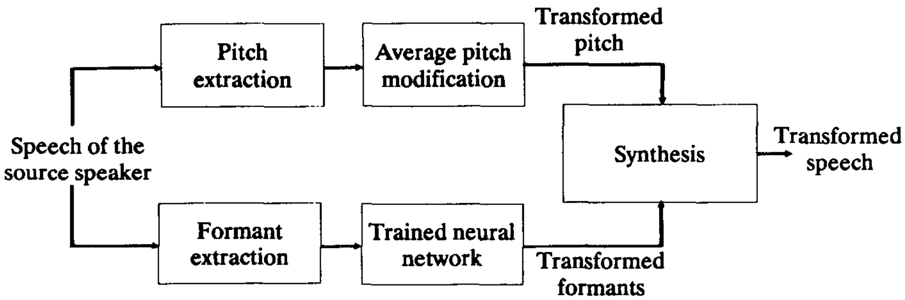


Fig. 6. Block diagram showing the transformation phase of the voice conversion system.

gain contour extracted from the speech of the source speaker was used directly without any modification for synthesis. Fant's model (Fant, 1986) was used to excite the formant synthesizer for voiced frames and random noise for the case of unvoiced frames. Transformed speech was obtained for three cases:

(a) Average pitch transformation: speech with original system and source characteristics modified by average pitch.

(b) Formant transition: speech with original source characteristics and the transformed formants.

(c) Average pitch and formant transformation: speech with original source characteristics modified by average pitch and transformed formants.

The third case which includes both formant transformation and pitch modification factor does indeed bring in the characteristics of the target speaker in the synthesized speech. However, several important speaker characteristics of the target speaker were not incorporated in the transformation, especially the glottal pulse shape and the prosodic features. We are currently exploring ways of incorporating these features in the transformation to produce a more natural sounding speech with the characteristics of the target speaker.

5. Summary and conclusion

In this paper we have described a general scheme for voice conversion. We have discussed the studies performed on interspeaker variation (gender differences) in the locations of formants and inherent pitch. We have demonstrated that a feedforward neural network trained using the backpropagation algorithm can capture a function which could transform the formants of the source speaker to that of the target speaker. Even though the network was trained with formants extracted from the steady voiced regions, it has faithfully transformed formant transitions as well. Pitch was modified using an average pitch modification factor. The transformed speech did possess characteristics of the female speaker (target speaker). The quality of the transformation can be improved by using glottal pulse shape transformation at the segmental level and pitch contour transformation at the prosodic level, in addition to the proposed formant and average pitch transformations.

References

M. Abe (1991), "A segment-based approach to voice conversion", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, pp. 765–768.

M. Abe, S. Nakamura, K. Shikano and H. Kuwabara (1988), "Voice conversion through vector quantization", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, pp. 655–658.

B.S. Atal and S.L. Hanauer (1971), "Speech analysis and synthesis by linear prediction of the speech wave", *J. Acoust. Soc. Amer.*, Vol. 50, No. 2, pp. 637–655.

D.G. Childers and C.K. Lee (1991), "Vocal quality factors: Analysis, synthesis and perception", *J. Acoust. Soc. Amer.*, Vol. 90, No. 5, pp. 2394–2410.

D.G. Childers, B. Yegnanarayana and K. Wu (1985), "Voice conversion: Factors responsible for quality", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, pp. 19.10.1–19.10.4.

D.G. Childers, K. Wu and D.M. Hicks (1987), "Factors in voice quality: Acoustic features related to gender", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, pp. 293–296.

D.G. Childers, K. Wu, D.M. Hicks and B. Yegnanarayana (1989), "Voice conversion", *Speech Communication*, Vol. 8, No. 2, pp. 147–158.

G. Fant (1986), "Glottal flow: Models and interaction", *J. of Phonetics*, Vol. 14, Nos. 3–4, pp. 393–399.

G. Fant, A. Kruckenberg and L. Nord (1991), "Prosodic and segmental speaker variations", *Speech Communication*, Vol. 10, Nos. 5–6, pp. 521–531.

K. Hornik, M. Stinchcombe and H. White (1989), "Multilayer networks are universal approximators", *Neural Networks*, Vol. 2, pp. 359–366.

D.H. Klatt and L.C. Klatt (1990), "Analysis synthesis and perception of voice quality variations among female and male speakers", *J. Acoust. Soc. Amer.*, Vol. 87, No. 2, pp. 820–857.

J.D. Markel (1972), "The SIFT algorithm for fundamental frequency estimation", *IEEE Trans. Audio Electroacoust.*, Vol. AU-20, pp. 367–377.

T.L. McClelland, D.E. Rumelhart and the PDP Research Group (1986), Parallel Distributed Processing (MIT press, Cambridge, MA).

H.A. Murthy and B. Yegnanarayana (1991), "Formant extraction from group delay function", *Speech Communication*, Vol. 10, No. 3, pp. 209–221.

L.R. Rabiner and B.H. Juang (1993), *Fundamentals of Speech Recognition* (Prentice Hall, Englewood Cliff, NJ).

M. Savic and I.H. Nam (1991), "Voice personality transformation", *Digital Signal Processing,* Vol. 4, pp. 107–110.

S. Seneff (1982), "System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction", *IEEE Trans. Acoust. Speech Signal Process.,* Vol. ASSP-30, No. 4, pp. 566–578.

H. Valbret, E. Moulines and J.P. Tubach (1992), "Voice transformation using PSOLA technique", *Speech Communication,* Vol. 11, Nos. 2–3, pp. 175–187.