ELSEVIER

Speech Communication 16 (1995) 153–164

SPEECH COMMUNICATION

Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt

Hideyuki Mizuno, Masanobu Abe *

NTT Human Interface Laboratories, 1-2356 Take, Yokosuka-shi, Kanagawa 238-03, Japan

Received 7 March 1994; revised 4 November 1994

Abstract

This article presents a new algorithm used in order to convert the speech of one speaker so that it sounds like that of another speaker. This algorithm flexibly converts voice quality using two major technical developments. Firstly, the modification of formant frequencies and spectral intensity using piecewise linear voice conversion rules. This enables the control of spectrum parameters in detail. The conversion rules are generated automatically for any pair of speakers. The reliability of the conversion rules is guaranteed because they are statistically generated using training data. Secondly, this algorithm provides the ability to produce speech with the desired formant structure by controlling formant frequencies, formant bandwidths and spectral intensity. Speech is iteratively modified in order to achieve the specified formant structure. Listening tests prove that the proposed algorithm converts speaker individuality while maintaining high speech quality.

Zusammenfassung

Dieser Artikel stellt einen neuen Algorithmus zur Umwandlung der Sprechweise einer Person in einer Weise vor, daß ihre Äußerungen wie die eines anderen Sprechers klingen. Dieser Algorithmus wandelt die Sprecheigenschaften nach flexiblen Prinzipien um und zeichnet sich durch Neuheit in zwei wichtigen Punkten aus. Davon ist einer die Umwandlung der Formant-Frequenzen und der spektralen Intensität durch Anwendung von Regeln für linear stückweise Stimmenumwandlung. Dadurch wird es möglich, auf einfache Weise Einzelheiten der spektraken Parameter zu steuern. Diese Umwandlungsregeln sind besonders zuverlässig, weil sie mit Hilfe von während der vorbereitenden Untersuchungen gewonnenen Daten erarbeitet werden. Dabei werden sie automatisch für jedes Paar von Sprechern erstellt. Die andere Neuheit besteht darin, daß zur Erzielung einer Sprechweise mit der gewünschten Formantstruktur nicht nur Formant-Frequenzen und Formant-Bandweiten, sondern auch die spektrale Intensität geregelt werden. Dabei werden die gesprochenen Worte iterativ zur Anpassung an die angezielte Formant-Struktur modifiziert. Auswertung von Hörversuchen hat gezeigt, daß es der dargestellte Algorithmus erlaubt, die Individualität von Sprechern unter Beibehaltung hoher Qualität umzuwandeln.

Résumé

Dans cet article, nous présentons un nouvel algorithme de transformation du timbre de la voix. Cette méthode présente deux particularités importantes. Premièrement, elle modifie les fréquences des formants et l'intensité

* Corresponding author. E-mail: ave@nttspch.ntt.jp; Fax: +81-468-591054; Tel.: +81-468-592457.

SSDI 0167-6393(94)00052-2

spectrale en utilisant des règles de conversion "linéaires par morceaux", permettant de contrôler les détails du spectre. Ces règles de conversion sont apprises par des méthodes statistiques pour chaque paire de locuteur, garantissant ainsi une grande fiabilité. La deuxième particularité est que l'algorithme prend en compte non seulement la fréquence et la bande passante des formants, mais aussi l'intensité spectrale (l'amplitude formantique). Ceci est obtenu au moyen d'une procédure itérative de modifications des pôles du filtre de synthèse. Des tests d'écoute ont démontré que cet algorithme transforme de façon efficace les caractéristiques individuelles du timbre du locuteur, tout en conservant une très bonne qualité de synthèse.

Keywords: Voice conversion; Formant frequency; Spectral intensity; Spectrum tilt; Piecewise linear; Listening test

1. Introduction

Voice conversion is a technique used in order to change or modify speaker individuality; i.e., speech uttered by one speaker is transformed in order to sound as if it had been articulated by another speaker. This technique lends itself to a variety of significant applications: providing speaker individuality to synthesis-by-rule speech, designing hearing aids appropriate for specific hearing problems, improving the intelligibility of abnormal speech uttered by a speaker who has speech organ problems, etc.

Voice conversion has only recently been implicated in research, however, some algorithms have already been proposed (Childers et al., 1989; Matsumoto and Inoue, 1990; Valbret et al., 1992).

To introduce this topic, the concept of treating voice conversion as a mapping problem on the spectrum space was proposed (Abe et al., 1988). The speaker's spectrum space was divided into subspaces by vector quantization or matrix quantization. Centroid vectors of the subspaces were the parameters that characterized a speaker's individuality, and the mapping function was defined as the correspondence of the centroid vectors. Using this concept, a voice conversion algorithm was proposed which produced efficient performance. One problem remained to be resolved: improving the quality of the converted speech so that it was as close to the quality of the original speech as possible. The use of vector quantization or matrix quantization was effective in generating conversion rules that were statistically guaranteed to be reasonable. However, this degraded the quality of the converted speech because the variability of the output signal pattern was restricted by the size of the codebook.

Formant frequency is one of the most important parameters characterizing speech. Formant frequency is also an important factor in specifying speaker characteristics, as illustrated in previous studies (Matsumoto et al., 1973; Itoh and Saito, 1982; Kuwabara and Ohgushi, 1987). Formant synthesizers can synthesize various voice qualities such as male speech, female speech, childish speech, etc. (Klatt, 1982). Judging from these facts, the introduction of formant frequency as the control parameter in voice conversion should prove beneficial. Moreover, based on the speech production theory, formant frequency clearly has physical significance (Fant, 1960; Franagan, 1972; Klatt and Klatt, 1990). Therefore, using formant frequency as a parameter not only controls parameters that are directly connected to the speech production process, but also calls upon the most effective research findings to date.

This paper proposes a new voice conversion algorithm that manipulates formant frequencies and spectrum tilt. This algorithm elaborates upon an earlier voice conversion concept which was proposed in order to generate formant conversion rules (Abe et al., 1988). The proposed algorithm implicates three major technical development. Firstly, the reliability of the formant conversion rules is statistically guaranteed because the rules are generated using training data from the speech of two speakers. Secondly, the formant frequencies are extracted in an easy and stable manner by referring to the formant frequency which is predetermined for each subspace. Thirdly, the exact formant conversion rules are available, because the speaker's spectrum space is divided into subspaces by vector quantization, and the formant conversion rules are determined for each subspace. Formant conversion

rules are produced in a piecewise linear manner. In this algorithm, spectral intensity is introduced as a new control parameter in order to specify the formant structure. Speech is iteratively modified in order to minimize the spectral intensity distance (MSD) (Mizuno et al., 1993). The quality of the converted speech is higher than that of the LPC vocoder, because the algorithm modifies the speech waveform. The size of the codebook does not restrict the range of the output signal pattern, rather only the number of formant conversion rules. Consequently, in terms of total voice conversion, the new voice conversion algorithm synthesizes speech of higher quality than the conventional algorithm.

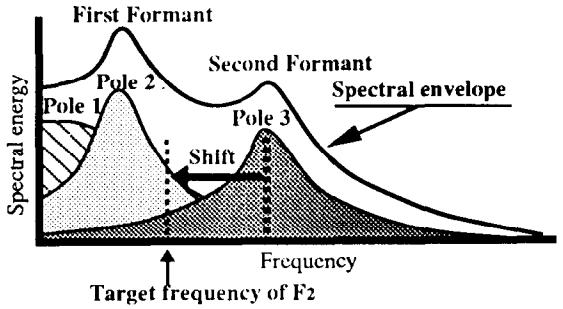
In Section 2, the new formant modification algorithm will be outlined. In Section 3, a new voice conversion algorithm will be presented. Section 4 will describe the listening tests that were conducted in order to evaluate the overall performance of the proposed voice conversion algorithm.

2. A new formant frequency modification algorithm

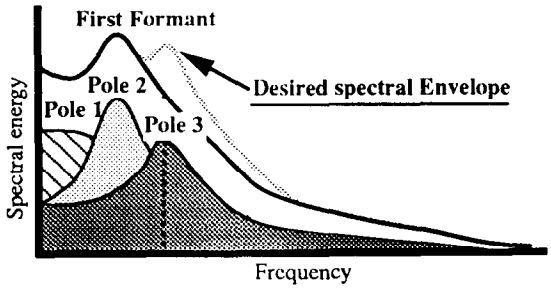
2.1. Outline of the algorithm

When speech signal is analyzed by an all-pole model, a speech formant is usually characterized by a cluster of poles. However, it is difficult to change a formant by shifting pole frequencies which are mainly related to the formant. The main problem is pole interaction. Fig. 1 shows an example of pole interaction.

In Fig. 1(a), the spectrum envelope is characterized by Pole 1, Pole 2 and Pole 3. Assume that Pole 2 and Pole 3 mainly characterize the 1st and 2nd formants, respectively. If only Pole 3 is modified in order to produce the desired spectrum of the lower 2nd formant, there is no maxima of the spectrum envelope at the target frequency of the 2nd formant, as illustrated in Fig. 1(b). In order to overcome this difficulty, a new parameter is introduced: spectral intensity, which is defined as the value of the LPC power spectrum density at a formant frequency. Moreover, in this paper, a formant is approximated by a pole which has the



(a) A spectral envelope with pole1, pole2 and pole3



(b)New spectral envelope with pole1, pole2 and shifted pole3

Fig. 1. Example for interaction.

narrowest bandwidth in a cluster of poles. Therefore, in the new algorithm, a formant is specified by a pole frequency, a pole bandwidth and a spectral intensity at the pole frequency, as shown in Fig. 2.

The specified formant structure is generated by two iterative processes. The first process generates the desirable formant structure. The second process generates a speech signal so that the formant structure of the synthesized speech is as close to the desired formant structure as possible. The formant modification procedure is applied to

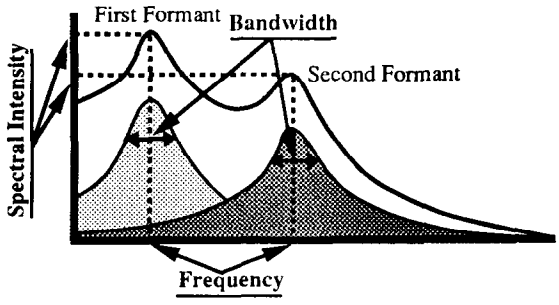


Fig. 2. Control parameters.

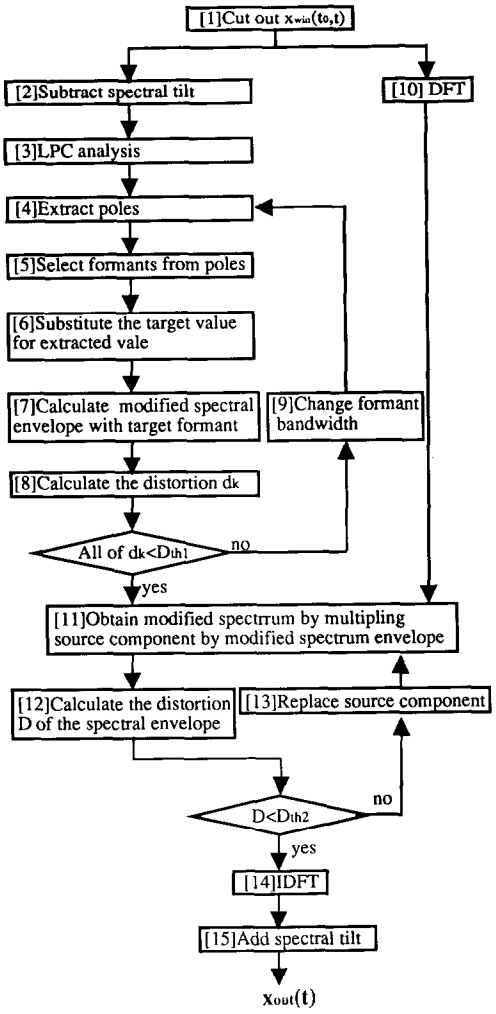


Fig. 3. Block diagram of formant frequency modification.

two pitch-periods of speech, and the successive analysis frames are shifted by one pitch-period. In the following section, the proposed algorithm is explained in detail, corresponding to Fig. 3. Detail parameters such as analysis order, frame length, threshold values, etc. will be shown in Section 4. The numbers which appear in the explanations refer to the block numbers shown in Fig. 3.

2.2. MSD algorithm procedure

[1] From input waveform $x_{in}(n)$, obtain waveform $x_{win}(t_0,n)$ at analysis time instant t_0 using a window function $w_h(n)$ of length N:

$$x_{win}(t_0,n) = w_h(n)x_{in}(n). \tag{1}$$

In the following, the notation t_0 is omitted just for convenience.

[2] Obtain tilt-corrected signal $x(n)$ by subtracting a spectral tilt of $x_{win}(n)$. Here, the spectrum tilt is approximated by the critical dumping filter $H(z)$ as follows:

$$H(z) = \frac{1}{(1+gz^{-1})^2} = \frac{1}{Q(z)}, \tag{2}$$

where g is a dumping coefficient. g is obtained in order to minimize the value given by the following equation:

$$\sum_{n=2}^{N-1} \left(x_{win}(n) - \left(1 - 2gx_{win}(n-1) - g^2 x_{win}(n-2)\right)\right)^2. \tag{3}$$

This results in the solution of the following polynominal:

$$g^3 + 3c_1 g^2 + (2+c_2)g + c_1 = 0, \tag{4}$$

where c_i is the i-th short-term autocorrelation coefficient of $x_{win}(n)$ normalized by power. Here, the real root of Eq. (4) is only used as the coefficient g. Finally, the tilt-corrected signal $x(n)$ is calculated in the following equation:

$$x(n) = \sum_{\tau=0}^{n} q(n-\tau)x_{win}(\tau), \tag{5}$$

where $q(n)$ is the impulse response of $Q(z)$.

[3] Calculate vocal tract transfer function $S(z)$ from $x(n)$ by LPC analysis. $S(z)$ is given by

$$S(z) = \frac{1}{1 + \sum_{i=1}^{p} \alpha_i z^{-i}}, \tag{6}$$

where α_i is the LPC coefficient of $x(n)$ ($i = 1,2,\ldots,p$; p is the LPC order).

[4] Calculate poles $\{z_i\}$ of $S(z)$, which are given as roots of the following polynomial:

$$1 + \sum_{i=1}^{p} \alpha_i z^{-i} = 0. \tag{7}$$

[5] Select poles Z_k that have low b_i/ω_i values. In the following procedures, a formant is approxi-

mated by Z_k ($k = 1,2,\ldots,m$; m is the number of formants) (Markel, 1972; Furui, 1989). Here, resonance frequency ω_i and its bandwidth b_i are given by

$$\omega_i = \arg(z_i), \quad b_i = \log|z_i|. \tag{8}$$

[6] In order to modify some of the extracted formant frequencies, substitute a desired (target) frequency ω_k^{tar} for $\arg(Z_k)$, of k-th formant Z_k. Here, the modified formant Z_k' is given by

$$Z_k' = \log|Z_k| e^{j\omega_k^{\text{tar}}}. \tag{9}$$

[7] Calculate a new vocal tract transfer function $S'(z)$ which is constructed by Z_k' and all other poles using Eq. (6). Finally, obtain the modified vocal transfer function $S^{\text{mod}}(z)$ by subtracting the spectral tilt of $S'(z)$ using the same method described in [2]. In this case, the autocorrelation coefficients, which are used in Eq. (4), are obtained by solving the following simultaneous equations:

$$\sum_{i=1}^{p} \alpha_i' c_{|i-j|}' = 0, \quad j = 1,2,\ldots,p, \tag{10}$$

where α_i' is the i-th LPC coefficient of $S'(z)$, and c_l' the l-th short-term autocorrelation coefficient. [8] Calculate the LPC spectrum envelope $F^{\text{mod}}(\omega)$ using the following equation:

$$F^{\text{mod}}(\omega) = \frac{1}{1 + \sum_{i=1}^{p} \alpha_i^{\text{mod}}\, e^{-j\omega i}}, \tag{11}$$

where α_i^{mod} is the i-th LPC coefficient of $S^{\text{mod}}(z)$. Calculate the spectral distortion d_k between the target spectral intensities I_k^{tar} and the spectral intensities $F^{\text{mod}}(\omega_k^{\text{tar}})$ using the following equation:

$$d_k = |F^{\text{mod}}(\omega_k^{\text{tar}})| - I_k^{\text{tar}}. \tag{12}$$

[9] If the distortion d_k is larger than the threshold D_{th1}, change the formant bandwidth $B_k = \log(Z_k)$ as follows:

$$B_k = \begin{cases} (1+\delta_k)B_k, & d_k > 0, \\ (1-\delta_k)B_k, & \text{elsewhere}, \end{cases} \tag{13}$$

$k = 1,\ldots,m$.

Here, δ_k (> 0) is the stepsize for bandwidth modification. When the sign of δ_k is turned over, d_k is divided by two. If any d_k are larger than D_{th1}, substitute $S^{\text{mod}}(z)$ for $S(z)$, and go to step [4].

[10] Obtain the short-time spectrum $W(w)$ by computing the DFT (Discrete Fourier Transform) of $x(n)$.

$$W(\omega) = \sum_{N=0}^{N-1} x(n) e^{-j\omega n}. \tag{14}$$

This can be calculated efficiently by resorting to the well-known FFT algorithm.

[11] The short-time spectrum is then split into the tilt-corrected envelope $F(\omega)$, and a source component $G(\omega)$, which is obtained by dividing the $W(\omega)$ by the tilt-corrected envelope $F(\omega)$.

$$G(\omega) = \frac{W(\omega)}{F(\omega)}. \tag{15}$$

Here, $F(\omega)$ is the normalized frequency notation of $S(z)$ and is given by

$$F(\omega) = \frac{1}{1 + \sum_{i=1}^{p} \alpha_i e^{-j\omega i}}, \tag{16}$$

where α_i is the LPC coefficient of $x(n)$ ($i = 1,2,\ldots,p$; p is the LPC order). Calculate the modified short-time spectrum envelope $W^{\text{mod}}(\omega)$ by multiplying the source component $G(\omega)$ by the modified spectrum envelope $F^{\text{mod}}(\omega)$ using Eq. (17).

$$W^{\text{mod}}(\omega) = G(\omega) F^{\text{mod}}(\omega). \tag{17}$$

[12] Calculate the LPC spectrum envelope $F'(w)$ from $W^{\text{mod}}(\omega)$. Calculate the spectral distance D between the envelope $F^{\text{mod}}(\omega)$ and $F'(\omega)$, using Eq. (18).

$$D = \sum_{k=1}^{m} \| F'(\omega)_k^{\text{tar}} | - | F^{\text{mod}}(\omega_k^{\text{tar}}) \|. \tag{18}$$

[13] If D is lager than the threshold D_{th2}, go to step [11], after substituting the $F'(\omega)$ for $F(\omega)$.

[14] Transform spectrum $W^{\text{mod}}(\omega)$ into signal $x^{\text{mod}}(n)$ using an inverse DFT. This can be calculated efficiently using inverse-FFT.

$$x^{\text{mod}}(n) = \frac{1}{N} \sum_{i=0}^{N-1} W^{\text{mod}}(\omega) e^{j(2\pi/N)ni}. \tag{19}$$

[15] Produce the final waveform $x_{out}(n)$ with the spectrum tilt by Eq. (20).

$$x_{out}(n) = \sum_{\tau=0}^{n} h(n-\tau)x^{mod}(\tau), \quad (20)$$

where $h(n)$ is the impulse response of filter $H(z)$ given by Eq. (2).

3. A new voice conversion algorithm

3.1. Outline

This algorithm consists of both on-line and off-line procedures. The off-line procedures generate the conversion rules for the formant frequencies and the spectrum tilt for each subspace obtained by vector quantization. In the on-line procedures, voice conversion is applied only to the voiced parts. All procedures are performed pitch-synchronously. Firstly, the input speech is vector-quantized in order to determine the appropriate subspace for the speech. Then, the formant frequencies are extracted using the reference formant frequencies. Finally, the formant frequencies and spectrum tilts are modified by the MSD algorithm according to the rules associated with each subspace.

3.2. Making voice conversion rules (off-line procedures)

The mapping codebooks are a pair of codebooks, where the code vectors of one codebook (original speaker's) have a one-to-one correspondence with the code vectors of the other codebook (target speaker's) (Abe et al., 1988). The following is a brief review of how to generate a mapping codebook.

[1] Using training utterances, generate a speaker-specific codebook for each speaker, say Speaker A and Speaker B, using by LBG algorithm (Linde et al., 1980). Here, both speakers must utter the same text as contained in the training utterances. Then, Speaker A's and Speaker B's utterances are vector quantized using his/her codebook.

[2] Using the code vector sequence for the same text utterance from the two speakers, the correspondence between the code vectors is determined by Dynamic Time Warping (Sakoe and Chiba, 1978).

[3] The code vector correspondences between the two speakers are stored as histograms for all training utterances.

[4] Using the histogram for each code vector of Speaker A as the weighting function, a mapping codebook from Speaker A to B is defined as a linear combination of Speaker B's vectors.

The following is an explanation of how to generate formant frequency conversion rules based on the mapping codebooks. Fig. 4 shows the LPC spectrum envelopes that were extracted from the corresponding code vectors of the mapping codebook. Table 1 shows the parameters used in order to generate the mapping codebook. The upper one is the target speaker's spectrum

Table 1
The analysis conditions to generate the mapping codebooks

Sampling frequency	12 kHz
LPC analysis	autocorrelation method
LPC analysis order	16
Frame shift length	3.0 msec
Frame length	30 msec
Window function	Hanning window
VQ distance measure	cepstrum distance
Code book parameter	autocorrelation coefficient
Speech material	isolated utterances of phonetically-balanced 216 words
Learning data size to generate codebooks	20,000 frames
Codebook size	256
Learning words to generate mapping codebooks	100

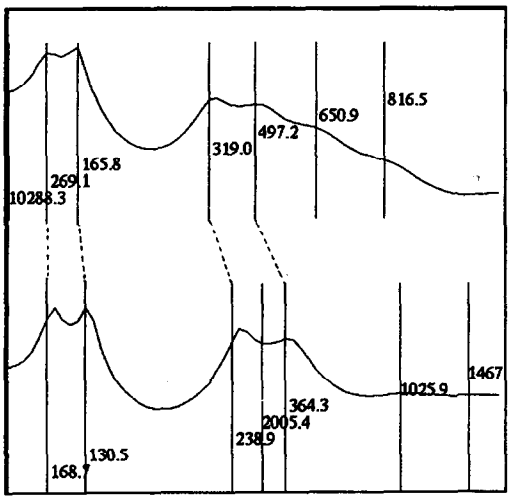


Fig. 4. The correspondence of the poles extracted from the mapping codebooks.

envelope and the lower one is that of the original speaker. In Fig. 4, pole frequencies are indicated by vertical lines, and the numbers indicate the bandwidth of each pole.

Taking into account the pole bandwidth by a direct visual inspection, certain poles are selected as formants (F_1, F_2, F_3 and F_4) for both the target and the original speaker. The formant correspondence is manually determined (the resulting correspondence is shown by the slanted lines in Fig. 4). At this stage, in order to correctly assign formant correspondence, this process is carried out manually. These experiences indicate that the formant correspondence is almost always successfully defined for vowel-like code vectors. Spectrum tilt is also obtained from the corresponding code vector. Spectrum tilt is approximated using the $H(z)$ given by Eq. (2). The original speaker's code vector, the formant frequency shift values and the spectrum tilt shift values are stored as the conversion rules. The extracted formant frequency is also stored as the reference formant frequency.

3.3. Voice conversion algorithm (on-line procedures)

Fig. 5 shows a block diagram of the proposed algorithm. In the following explanation, the numbers refer to the block numbers cited in Fig. 5.

[1] Input speech uttered by the original speaker is vector quantized using the original speaker's codebook.

[2] Outputs of [1] are smoothed by using the code vector buffer. This buffer includes the current code vector and the three previous code vectors. The code vector which occurs most frequently in the buffer is selected as an output code vector. This smoothing technique results in the accurate extraction of formant frequencies in [3] and the accurate generation of target formant frequencies in [4].

[3] Depending upon the output code vector, reference formant frequencies are set using the stored data generated in off-line procedures.

[4] As described in Section 2.2, LPC poles and spectrum tilts are calculated from the input

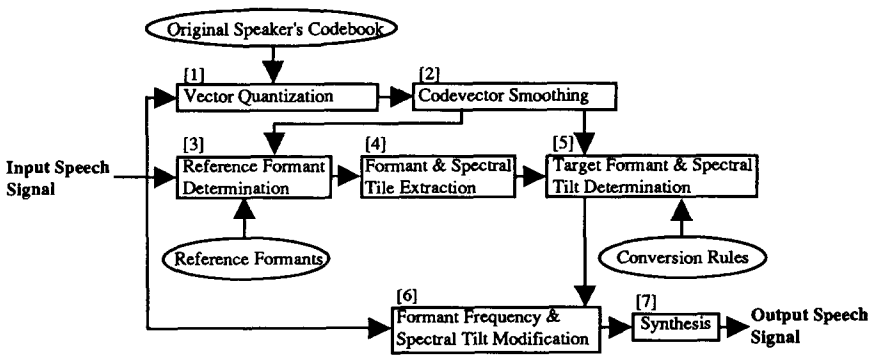


Fig. 5. Block diagram of voice conversion algorithm.

speech. If a pole has a lower b_i/w_i value, and the pole frequency is the nearest to the reference formant frequency, the LPC poles are selected as formant.

[5] The target formant frequency is the sum of the extracted formant frequency and the formant shift value which is determined by the associated code vector. The target spectral tilt is obtained in the same way. The target spectral intensity is the value of the input spectral intensity.

[6] The formant frequencies and spectrum tilt are modified by the MSD in order to yield the target values.

[7] The speech signal is synthesized by pitch-synchronous overlap addition.

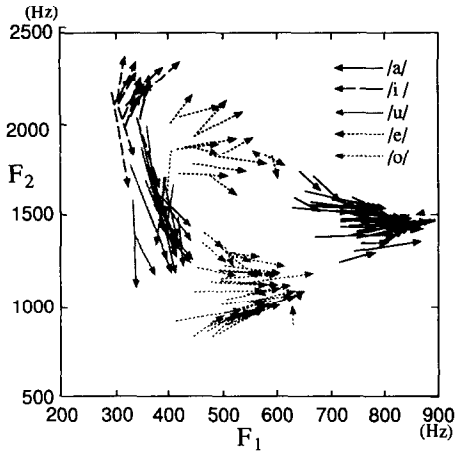


Fig. 6. Conversion rules of F_1 and F_2. The arrows show the direction and magnitude of formant transformation from original to target.

4. Performance evaluation

4.1. Preliminary evaluation experiments

In order to evaluate the proposed algorithm, the accuracy of formant tracking was first examined. Table 2 shows the experimental conditions. F_0 values were modified using the PSOLA technique (Hamon et al., 1989) in order to match the average F_0 of the target speaker. An evaluation test was conducted using 5 continuously uttered sentences. Table 3 illustrates the results. When the automatically determined pole is not equal to the manually selected pole, it is considered to be an error. Judging from the results, the proposed algorithm shows very good performance, in terms of F_1 and F_2 tracking for vowels. In other formants, extraction errors occurred mainly when the formant frequency had low power in the spectrum domain. An informal listening test confirmed that automatic formant tracking, and manual formant tracking, yielded converted speech of equal quality.

Fig. 6 shows the formant frequency shift rules on the F_1–F_2 plane for two male speakers. The arrows indicate formant shifts from the original speaker to the target speaker. Formant frequency shift rules are precisely described for each vowel, and the streams of the shift directions indicate quite reasonable conversion.

Fig. 7 illustrates one example of the speech spectrograms. In the converted speech (Fig. 7(b)), it is clearly evident that F_3 has moved downward from the original speech (Fig. 7(a)) to F_3 of the target speech (Fig. 7(c)). Moreover, the formant frequency shift is quite smooth.

Table 2
The analysis conditions for MSD

Sampling frequency	12 kHz
LPC analysis	autocorrelation method
LPC order	16
Window function	Hanning window
Frame length	2 pitch period
The number of formant	4(F_1–F_4)
The threshold of b_i/ω_i	0.15
Initial stepsize δ_k for bandwidth modification	0.2
FFT points	1024
Threshold	D_{th1} = 0.2 (dB) D_{th2} = 1.0 (dB)

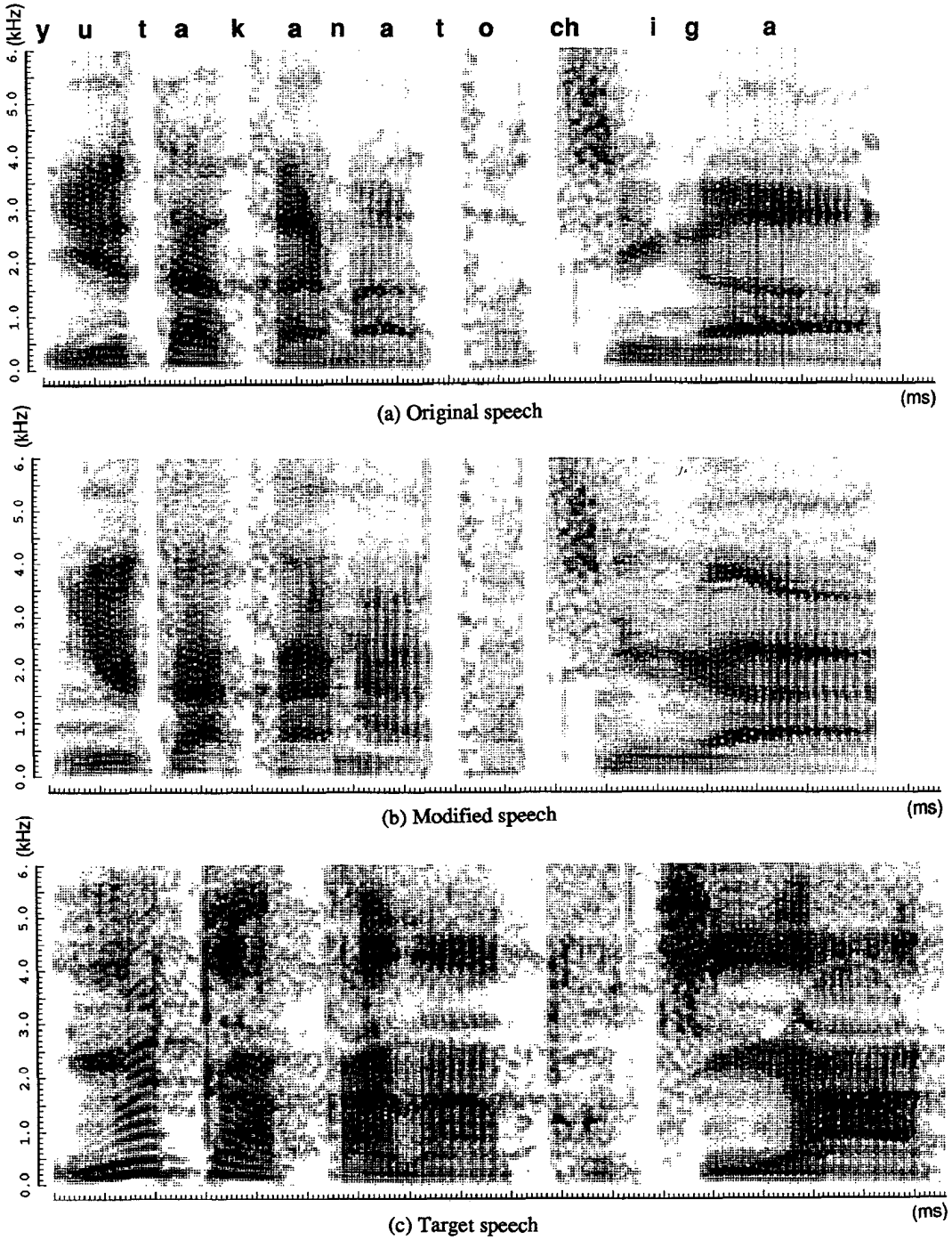


Fig. 7. Comparison of speech spectrograms.

Table 3
The error rate of formant extraction [error rate (%); error number/total number]

	Vowel	Voiced consonant
F_1	0.16 (2/948)	0 (0/284)
F_2	1.27 (12/948)	16.55 (47/284)
F_3	13.40 (127/948)	19.72 (56/284)
Average	4.96 (141/2844)	12.09 (103/852)

4.2. Evaluation by listening tests

Three kinds of listening tests were conducted, comprising a total of ten listeners. Four speakers provided speech material; three were men (man1, man2 and man3) and one was an 11-year-old boy. In all tests, man1's voice was converted in order to match those of the others.

Table 4 illustrates the spectrum distortion and average F_0 differences between man1 and the others. In terms of speech quality, man3 was closest to man1, while the 11-year-old boy was farthest. Fig. 8 shows the formant frequency shift values on the F_1–F_2 plane for each speaker pair. For the man1–boy pair (Fig. 8(a)), the shift values and shift directions are very different from those of the other pairs. This is mainly due to the fact that adults and children have speech organs which are quite different in size. In terms of man1–man2 (Fig. 8(b)) and man1–man3 (Fig. 8(c)), the shift directions are similar, but the shift values of man1–man2 are larger than those for man1–man3.

4.2.1. XAB test

The first listening test was designed in order to evaluate the speaker individuality conversion accuracy using the XAB method. Stimuli A and B were either the original speaker's speech (man1),

Table 4
The average F_0 difference and cepstrum distortion (CD) at mapping codebook generation

Target speaker	Difference of F_0 (Hz)	CD
Boy	93.3	0.84
Man2	40.4	0.61
Man3	23.5	0.56

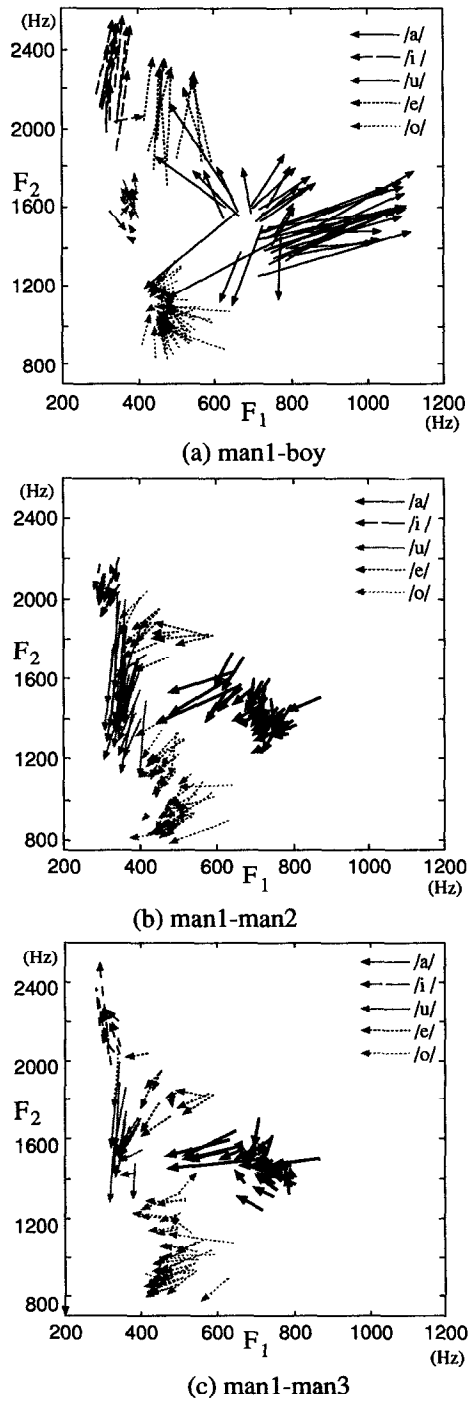


Fig. 8. Conversion rules of F_1 and F_2. The arrows show the direction and magnitude of formant transformation from original (man1) to target (boy, man2 and man3).

or the target speaker's speech (man2, man3 and/or boy). Stimulus X was either the converted speech using the codebook mapping method, the converted speech using the proposed algorithm, the original speaker's speech, or the target speaker's speech. Three different words were used for the conversions, and each triad was a combination of two different words. Listeners were asked to select either A or B as being the closest to X.

Table 5 shows speaker identification rates. The voice conversion performance depended upon the distance between the original speaker and the target speaker. For the man1–boy pair, the converted speech was clearly identified, while the converted speech was indistinct for man1–man3. The trend in identification rates concurs with the trend in the similarity of speech quality shown in Table 4.

When comparing the codebook mapping method and the proposed algorithm, no difference is noted in voice conversion performance. This is quite logical because, in terms of speaker individuality, both algorithms use the same information. The conversion rules of the proposed algorithm were generated from the mapping codebook.

4.2.2. Preference test

This test was designed in order to evaluate the speech quality of both the codebook mapping method and the proposed algorithm. Each test pair consisted of the converted speech produced by the codebook mapping method, and the converted speech which was output by the proposed algorithm. The words were the same as those used in the XAB test. Listeners were asked to indicate their preference for each pair.

Fig. 9 shows the experimental results. These results confirm that the proposed algorithm has significantly better performance than conventional methods. This is mainly due to the fact that the proposed algorithm directly modifies the spectrum of the waveform.

4.2.3. Opinion test

In order to evaluate the overall performance of the proposed algorithm, an opinion test was

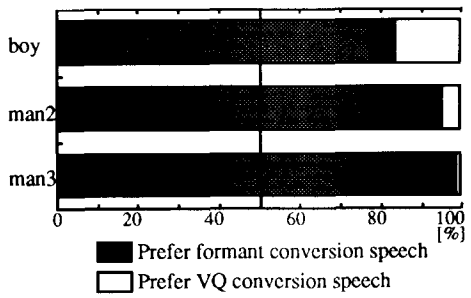
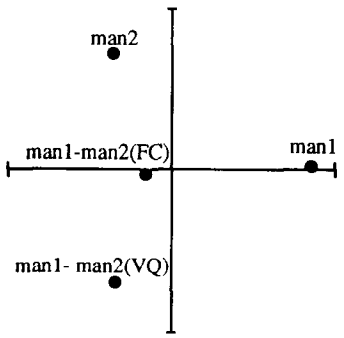


Fig. 9. Preference score.

carried out. Each speech pair consisted of two different words from 4 different groups. The groups were man1's speech, man2's speech, the speech converted by the proposed algorithm and the speech converted by the codebook mapping method (man1's to man2's).

Listeners were asked to rate the similarity of each pair in five categories, ranging from "similar" to "dissimilar". Hayashi's fourth method of quantification was applied to the experimental data obtained by the listening test. This method places a sample in a space according to the similarities between the samples (Hayashi, 1985).

The representation of the results on a two-dimensional illustration is shown in Fig. 10. This figure shows the relative similarity-distance between stimuli. The contribution factors of the first and the second axis were 0.74 and 0.26,



man1: original man1's speech
man2: original man2's speech
man1-man2(VQ): VQ conversion
man1-man2(FC) :formant conversion

Fig. 10. Distribution of psychological distances for man1-to-man2 voice conversion.

Table 5
XAB test results

Target speaker	Correct response (%)	
	VQ	Formant
Boy	92	88
Man2	62	60
Man3	48	35

respectively. The speech converted by the proposed algorithm is much closer to that of the target speaker than the speech converted by the mapping codebook method. One of the reasons is that, as explained in Section 4.2.2, the proposed algorithm synthesizes high quality speech.

5. Conclusion

A new voice conversion algorithm has been proposed. This algorithm is characterized by the use of piecewise linear conversion rules in order to precisely determine the target formant frequencies and spectrum tilt. The algorithm is also characterized by the iterative modification method that yields the target formant frequencies. Preliminary tests have confirmed that the algorithm produces reasonable formant frequency conversion rules and smoothly modifies the formant frequencies. In order to evaluate the performance of the proposed algorithm, three kinds of listening tests were carried out. The experimental results showed that, when compared to the conventional algorithms, the proposed algorithm enables the conversion of speaker individuality, while maintaining high speech quality. There are two reasons: firstly, the algorithm directly modifies speech waveforms in the spectrum domain; and secondly, the size of codebook does not restrict the range of the output signal patterns, rather only the number of formant conversion rules.

Acknowledgements

The authors wish to thank Dr. N. Kitawaki, Director of the Speech and Acoustic Laboratory, and Dr. N. Sugamura, Leader of Speech Processing Group, for their encouragement during this contribution.

References

M. Abe, S. Nakamura, K. Shikano and H. Kuwabara (1988), "Voice conversion through vector quantization", *Internat. Conf. Acoust. Speech Signal Process.-'88,* pp. 565–568.

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 (1960), *Acoustic Theory of Speech Production* (Mouton, The Hague).

J.L. Franagan (1972), *Speech Analysis Synthesis and Perception,* 2nd Edition (Springer, Berlin).

S. Furui (1989), *Digital Speech Processing, Synthesis, and Recognition* (Dekker, New York), p. 97.

C. Hamon, E. Moulines and F. Chaarpentier (1989), "A diphone synthesis system based on time-domain prosodic modification of speech", *Internat. Conf. Acoust. Speech Signal Process.-'89,* pp. 238–241.

C. Hayashi (1985), "Recent theoretical and methodological developments in multidimensional scaling and its related method in Japan", *Behaviormetrika,* No. 18, 1095.

K. Itoh and S. Saito (1982), "Effects of acoustical feature parameters of speech and its application of talker recognition", *Trans. IEICE Japan,* Vol. J65-A, pp. 101–108.

D.H. Klatt (1982), "The Klattalk text-to-speech system", *Internat. Conf. Acoust. Speech Signal Process.-'82,* pp. 1589–1592.

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

H. Kuwabara and K. Ohgushi (1987), "Contributions of vocal tract resonant frequencies and bandwidths to the personal perception of speech", *Acoustica,* Vol. 63, pp. 120–128.

Y. Linde, A. Buzo and R.M. Gray (1980), "An algorithm for vector quantizer design", *IEEE Trans. Commun.,* Vol. COM-28, No. 1, pp. 84–95.

J.D. Markel (1972), "Digital inverse filtering – A new tool for formant trajectory estimation", *IEEE Trans. Audio Electroacoust.,* Vol. AU-20, No. 2, pp. 129–137.

H. Matsumoto and H. Inoue (1990), "A minimum distortion spectral mapping applied to voice quality conversion", *ICSLP'90,* pp. 161–164.

H. Matsumoto, S. Hiki, T. Sone and T. Nimura (1973), "Multidimensional representation of personal quality of vowels and its acoustical correlates", *IEEE Trans. AU,* Vol. AU-21, pp. 428–436.

H. Mizuno, M. Abe and T. Hirokawa (1993), "Waveform-based speech synthesis approach with a formant frequency modification", *Internat. Conf. Acoust. Speech Signal Process.-'93,* pp. 195–198.

H. Sakoe and S. Chiba (1978), "Dynamic programming algorithm optimization for spoken word recognition", *IEEE Trans. Acoust. Speech Signal Process.,* Vol. ASSP-26, No. 1, pp. 43–49.

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