Available online at www.sciencedirect.com
SCIENCE DIRECT
Speech Communication 48 (2006) 207–219
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Spectral and prosodic transformations of hearing-impaired Mandarin speech

Cheng-Lung Lee [a], Wen-Whei Chang [a,*], Yuan-Chuan Chiang [b]

[a] Department of Communications Engineering, National Chiao-Tung University, Hsinchu 300, Taiwan, ROC
[b] Department of Special Education, National Hsinchu Teachers College, Hsinchu, Taiwan, ROC

Received 21 January 2005; received in revised form 27 July 2005; accepted 17 August 2005

Abstract

This paper studies the combined use of spectral and prosodic conversions to enhance the hearing-impaired Mandarin speech. The analysis-synthesis system is based on a sinusoidal representation of the speech production mechanism. By taking advantage of the tone structure in Mandarin speech, pitch contours are orthogonally transformed and applied within the sinusoidal framework to perform pitch modification. Also proposed is a time-scale modification algorithm that finds accurate alignments between hearing-impaired and normal utterances. Using the alignments, spectral conversion is performed on subsyllabic acoustic units by a continuous probabilistic transform based on a Gaussian mixture model. Results of perceptual evaluation indicate that the proposed system greatly improves the intelligibility and the naturalness of hearing-impaired Mandarin speech.

Keywords: Voice conversion; Prosodic modification; Spectral conversion; Hearing-impaired speaker; Sinusoidal model

1. Introduction

Speech communication by profoundly hearing-impaired individuals suffers not only from the fact that they cannot hear other people's utterances, but also from the poor quality of their own productions. Due to the lack of adequate auditory feedback, the hearing-impaired speakers produce speech with segmental and suprasegmental errors (Hochberg et al., 1983). It is common to hear their speech flawed by misarticulated phonemes, with varying degrees of severity associated with their hearing thresholds (Monsen, 1978; McGarr and Harris, 1983). Their speech intelligibility is further affected by abnormal control over phoneme duration and pitch variations. Specifically, the duration of vowels, glides, and nasals were longer while the duration of fricatives, affricates, and plosives were shorter than in normal speech, and the pitch contour over individual syllables is either too varied or too monotonous. Their intonation also shows limited pitch variation, erratic pitch fluctuations, and inappropriate average F0 (Osberger and Levitt, 1979). This motivates our research into trying to devise a voice conversion system that modifies the speech of a hearing-impaired (source) speaker to

* Corresponding author. Tel.: +886 3 5731826; fax: +886 3 5710116.
E-mail address: wwchang@cc.nctu.edu.tw (W.-W. Chang).

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.08.001

be perceived as if it was uttered by a normal (target) speaker. The technique of voice conversion has applications in text-to-speech synthesis (Kain and Macon, 1998) and improving the quality of alaryngeal speech (Bi and Qi, 1997). Most current systems (Abe et al., 1988; Stylianou et al., 1998) concentrate on the spectral envelope transformation while the conversion of prosodic features is essentially obtained through a simple normalization of the average pitch. Such systems may lead to an unsatisfactory speech conversion quality in cases of tonal languages, such as Chinese, which uses lexical tones to distinguish meanings of syllables that have the same phonetic compositions. In view of the important roles of prosody in Mandarin speech perception, further enhancement is expected by better modelling of pitch contour dynamics and by additionally incorporating prosodic transformation into the voice conversion system.

The key to solving the problem of voice conversion lies in the detection and exploitation of characteristic features that distinguish the impaired speech from the normal speech (Ohde and Sharf, 1992). To proceed with this, we found the phonological structure of Chinese language could be used to advantage in the search for the basic speech units for prosodic and spectral manipulations. Mandarin Chinese is a tonal language in which each syllable, with few exceptions, represents a morpheme (Lee, 1997). Traditional descriptions of the Chinese syllable structure divide syllables into combinations of initials and finals rather than into individual phonemes. An initial is the consonant onset of a syllable, while a final comprises a vowel or diphthong but includes a possible medial or nasal ending. Depending on the manner of articulation, initial consonants can be further categorized into five phonetic classes including fricatives, affricates, stops, nasals, and glides. To convey different lexical meanings, each syllable can be pronounced with four basic tones; namely, the high-level tone (tone 1), the rising tone (tone 2), the falling-rising tone (tone 3), and the falling tone (tone 4), which are acoustically correlated with different fundamental frequency (F0) contours and use duration and intensity of the vowel nucleus to provide secondary information. Recent perceptual work on Chinese deaf speech (Chang, 2000; Lin and Huang, 1997) has shown that speakers with greater than moderate degrees of losses ($\geqslant 50$ dB HL bilaterally) were perceived with an average accuracy of 31% in phoneme production, and further, that the most errors in the consonants were affricates and fricatives. This finding may have more serious implications for Mandarin than for other languages as these two phonetic classes make up more than half of the consonants in Mandarin Chinese. Moreover, since most of them are palatal or produced without apparent visual cues, they are difficult to correct through speech training. In tone production, their accuracy only reached an average of 54%, with most errors involving confusions between tones 1 and 4, tones 1 and 2, and tones 2 and 3. The results also showed that tones produced by speakers with profound losses were only half as likely to be judged correct as those produced by speakers with less loss. Again, as tones are produced by phonatory, rather than articulatory control, they are almost impossible to correct through non-instrumental-based speech therapy. In view of the prevalence of the problems in hearing-impaired Mandarin speech, we propose a subsyllable-based approach to voice conversion that takes into consideration both the prosodic and the spectral characteristics. The target application of our technique will be in computer-assisted language learning. As the three aspects of speech, i.e., spectrum, duration, and pitch, are manipulated independently, the hearing-impaired user can during a certain stage of learning choose to convert one aspect and use his/her own modified utterance as the target for focused correction. We believe the user can better perceive the conversion effect through auditory comparison of that aspect while errors in the other two aspects are temporarily ignored. This feature will thus guide the user going through a learning process much simpler than simply giving him/her an example spoken by a normal speaker with changes in all three aspects conglomerated.

2. System implementation

The general approach to voice conversion consists of first analyzing the input speech to obtain characteristic features, then applying the desired transformations to these features, and synthesizing the corresponding signal. Essentially, the production of sound can be described as the output of passing a glottal excitation signal through a linear system representing the characteristics of the vocal tract. To track the nonstationary evolution of characteristic features, both the spectral and prosodic manipulations will be performed on a frame-by-frame basis. In this work, speech signals were sampled at 11 kHz and analyzed using a 46.4 ms

Hamming windows with a 13.6 ms frame shift. Therefore, the analysis frame interval Q was fixed at 13.6 ms. For the speech on the mth frame, the vocal tract system function can be described in terms of its amplitude function $M(w; m)$ and phase function $\Phi(w; m)$. Usually the excitation signal is represented as a periodic train during voiced speech, and is represented as a noise-like signal during unvoiced speech. An alternative approach (McAulay and Quatieri, 1995) is to represent the excitation signal by a sum of $K(m)$ sine waves, each of which is associated with the frequency $w_k(m)$ and the phase $\Omega_k(m)$. Passing this excitation signal through the vocal tract system results in a sinusoidal representation of speech production. As noted elsewhere (Quatieri and McAulay, 1992), this sinusoidal framework allows flexible manipulation of speech parameters such as pitch and speaking rate while maintaining high speech quality.

A block diagram of the proposed voice conversion system is shown in Fig. 1. The system has five major components: speech analysis, spectral conversion, pitch modification, time-scale modification, and speech synthesis. The analysis begins by estimating from the Fourier transform of input speech the pitch period $P_0(m)$, the voicing probability $P_v(m)$, and the system amplitude function $M(w; m)$. The voicing probability will be used to control the harmonic spectrum cutoff frequency, $w_c(m) = \pi P_v(m)$, below which voiced speech was synthesized and above which unvoiced speech was synthesized. The second step in the analysis is to represent the system amplitude function $M(w; m)$ in terms of a set of cepstral coefficients $\{c_l(m)\}_{l=0}^{24}$. The main attraction of cepstral representation is that it exploits the minimum-phase model, where the log-magnitude and phase of the vocal tract system function can be uniquely related in terms of the Hilbert transform (Oppenhein and Schafer, 1989). A more comprehensive discussion of the sine-wave speech model and the corresponding analysis-synthesis system can be found in (McAulay and Quatieri, 1995).

The main part of the modification procedure involves the manipulation of functions which describe the amplitude and phase of the excitation and vocal tract system contributions to each sine-wave component. The effectiveness of voice conversion depends on a successful modification of prosodic features, especially of the time-scale and the pitch-scale. With reference to the sinusoidal framework, speech parameters included in the prosodic conversion are $P_0(m)$, $P_v(m)$, and the synthesis frame interval. The time-scale modification involves scaling the synthesis frame of original duration Q by a factor of $\rho(m)$, i.e., $Q'(m) = \rho(m)Q$. The pitch modification can be viewed as a transformation which, when applied to the pitch period $P_0(m)$, yields the new pitch period $P_0'(m)$, with an associated change in the F0 as $w_0'(m) = 2\pi/P_0'(m)$. It is worth noting also that the change in pitch period also corresponds to modification of the sine-wave frequencies $w_k'(m)$ and the excitation phases $\Omega_k'(m)$ used in the reconstruction. Below the cutoff frequency the sine-wave frequencies are harmonically related as $w_k'(m) = kw_0'(m)$, whereas above the cutoff frequency $w_k'(m) = k^* w_0'(m) + w_u$, where k^* is the largest value of k for which $k^* w_0' \leqslant w_c(m)$, and where w_u is the unvoiced F0 corresponding to 100 Hz. A two-step procedure is used in estimating the excitation phase $\Omega_k'(m)$ of

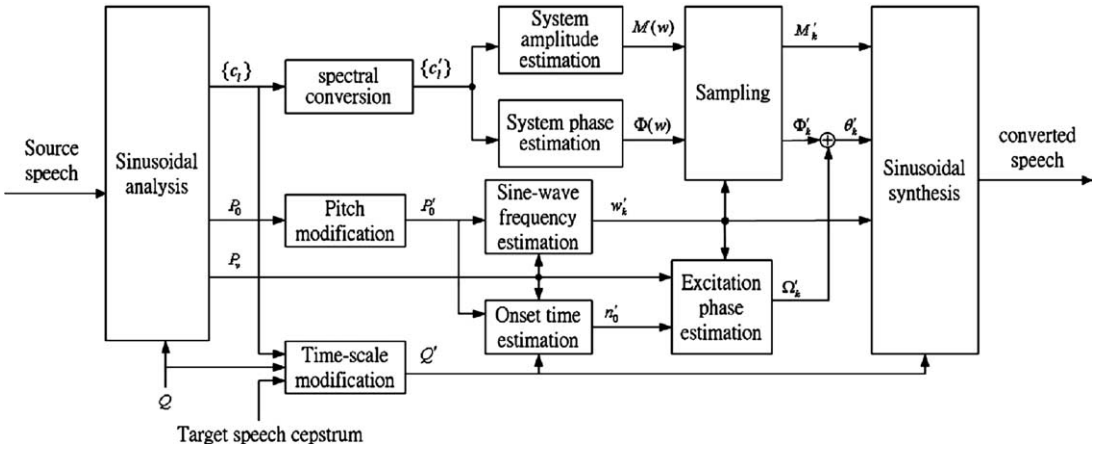


Fig. 1. Block diagram of the voice conversion system.

the kth sine wave. The first step is to obtain the onset time $n'_0(m)$ relative to both the new pitch period $P'_0(m)$ and the new frame interval $Q'(m)$. This is done by accumulating a succession of pitch periods until a pitch pulse crosses the center of the mth frame. The location of this pulse is the onset time $n'_0(m)$ at which sine waves are in phase. The second step is to compute the excitation phase as follows:

$$\Omega'_k(m) = -n'_0(m)w'_k(m) + \epsilon'_k(m), \tag{1}$$

where the unvoiced phase component $\epsilon'_k(m)$ is zero for the case of $w'_k(m) \leqslant w_c(m)$ and is made random on $[-\pi, \pi]$ for the case of $w'_k(m) > w_c(m)$.

In addition to prosodic conversion, the technique of spectral conversion is also needed to modify the articulation-related parameters of speech. The problem with the spectral conversion lies with the corresponding modification of the vocal tract system function. Thus there is a need to estimate the amplitude function $M'(w; m)$ and the phase function $\Phi'(w; m)$ of the vocal tract system. If it is assumed that the vocal tract system function is minimum phase (Oppenhein and Schafer, 1989), the log-magnitude and phase functions form a Hilbert transform pair and hence can be estimated from a set of new cepstral coefficients $\{c'_l(m)\}_{l=0}^{24}$. The system amplitudes $M'_k(m)$ and phases $\Phi'_k(m)$ are then given by samples of their respective functions at the new frequencies $w'_k(m)$, i.e., $M'_k(m) = M'(w'_k; m)$ and $\Phi'_k(m) = \Phi'(w'_k; m)$. Finally, in the synthesizer the system amplitudes are linearly interpolated over two consecutive frames. Also, the excitation and system phases are summed and the resulting sine-wave phases, $\theta'_k(m) = \Omega'_k(m) + \Phi'_k(m)$, are interpolated using the cubic polynomial interpolator. The final synthetic speech waveform on the mth frame is given by

$$s(n) = \sum_{k=1}^{K(m)} M'_k(m)\cos[nw'_k(m) + \theta'_k(m)],$$
$$t_m \leqslant n \leqslant t_{m+1} - 1, \tag{2}$$

where $t_m = \sum_{i=1}^{m-1} Q'(i)$ denotes the starting time of the current synthesis frame.

3. Time-scale modification

As stated earlier (Osberger and Levitt, 1979), the speech of the hearing-impaired speakers contains numerous timing errors, including a lower speaking rate, insertion of long pauses, and failure to modify segment duration as a function of phonetic environment. In Fig. 2 the mean phoneme durations produced by the hearing impaired were plotted against those produced by the normal speakers. Data were collected from two normal speakers (one male and one female) and three hearing-impaired speakers (one male and two females), all aged 15. The phonemes tested were five fricatives, six affricates, and three vowels. It can be seen that the mean duration ratios of impaired-to-normal utterances were quite different for different phonemes and that vowels, as a group, stayed much in line with the normal production than the two consonant groups, with the mean ratios for vowels,

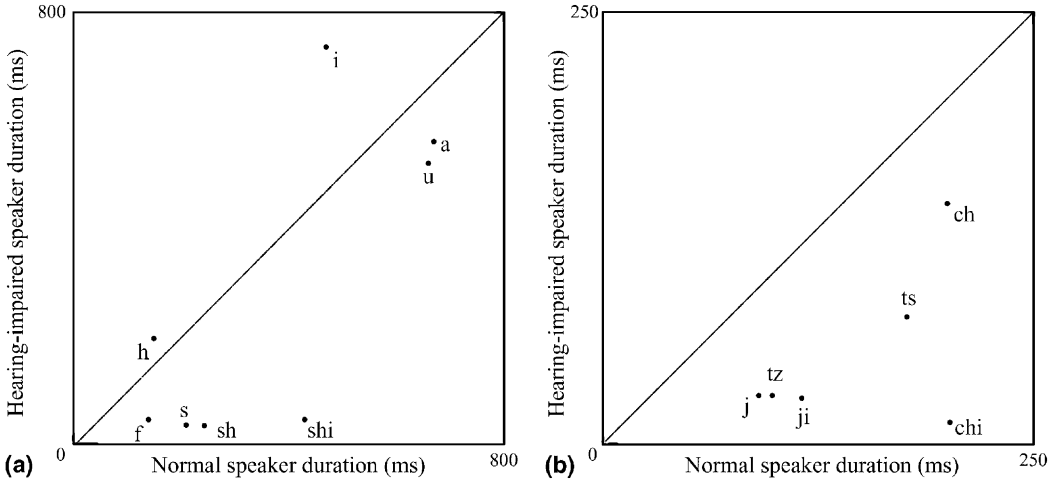


Fig. 2. Phoneme duration statistics for: (a) vowels and fricatives and (b) affricates.

fricatives, and affricates being 1.12, 0.4, and 0.34, respectively. All consonants, with the exception of /h/, of the hearing impaired were shorter, as indicated by their uniform appearances on the lower half of the graph. Our perceptual judgment showed that this shortening that could measure 10 to 1 (as seen in /sh/ and /shi/) was the result of substituting the two consonant classes with stops. These deviations can be corrected only when the system knows what the speaker meant to say. However, in analyzing the articulation errors made by the hearing- impaired speakers, one usually finds a lack of one-to-one substitution pattern between the error and the target, making it useless for incorporating automatic speech recognition into our system. Instead, our application made use of text prompt on the computer screen to elicit targets from the user during testing, a practice that has been widely used by commercialized software for speech training/correction.

For the converted speech to carry the naturalness of human speech, the duration of individual phonemes needs to match those found in the natural speech. This can be done by modifying the interval of each synthesis frame by a time-varying factor $\rho(m)$ in a way of $Q'(m) = \rho(m)Q$. The case $\rho(m) > 1$ corresponds to a time-scale expansion, while the case $\rho(m) < 1$ corresponds to a time-scale compression. The next step is to determine the time-scaling factor $\rho(m)$ based on spectral representations of the same syllable uttered by the source and target speakers. In describing the source speaker's spectral envelope, cepstral coefficients are measured frame by frame and are of the following form: $\boldsymbol{X} = \{\boldsymbol{x}(m_x), m_x = 1, 2, \ldots, T_x\}$, where T_x is the syllable duration in frames. Similarly, $\boldsymbol{Y} = \{\boldsymbol{y}(m_y), m_y = 1, 2, \ldots, T_y\}$ is the sequence of T_y cepstral vectors representing the target speaker's spectral envelope. Acoustic analysis of Mandarin hearing-impaired speech has indicated that unvoiced sound such as consonants may not be subjected to the same scaling as the vowels. Thus for time-scaling of speech, different approaches should be applied in the time-intervals where the frames corresponding to both speakers were marked as Mandarin initials or finals. The boundary between the initial and final parts of an isolated syllable is relatively easy to detect by a voiced/unvoiced decision based on the voicing probability P_v. Let B_x and B_y represent the starting frame for the final subsyllables in the source and target utterances, respectively. For constituent frames of the initial consonant, a linear time normalization was applied with a fixed factor $\rho = (B_y - 1)/(B_x - 1)$. With regards to the final subsyllables, two sets of paired cepstral vectors, $\{\boldsymbol{x}(m_x), B_x \leqslant m_x \leqslant T_x\}$ and $\{\boldsymbol{y}(m_y), B_y \leqslant m_y \leqslant T_y\}$, were time aligned using the procedure of dynamic time warping (DTW) (Rabiner and Juang, 1993). Usually the problem of DTW is formulated as a path finding problem over a finite range of grid points (m_x, m_y). The basic strategy applied here is to interpret the slope of the DTW path as a time-scaling function, which indicates on a frame-by-frame basis how much to shorten or lengthen each frame of the source utterance in order to reproduce the same duration as in target utterance.

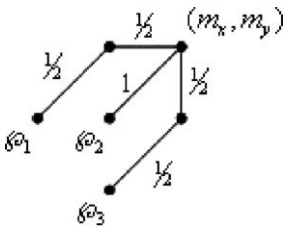
The DTW aims to align two utterances with a path through a matrix of similarity distances that minimizes the sum of the distances. We begin by defining a partial accumulated distance $D_A(m_x, m_y)$, representing the accumulated distance along the best path from the point (B_x, B_y) to the point (m_x, m_y). For an efficient implementation, a dynamic programming recursion is applied to compute $D_A(m_x, m_y)$ for all local paths that reach (m_x, m_y) in exactly one step from an intermediate point (m'_x, m'_y) using a set of local path constraints. Table 1 summarizes the local constraints and slope weights for three local paths, $\wp_1$, $\wp_2$, and $\wp_3$, chosen for the implementation. The local distance $d(m_x, m_y)$ between the time-aligned pairs of cepstral vectors is defined by a squared Euclidean distance. We summarize the dynamic programming implementation for finding the time-scaling factor at every frame of a final subsyllable as follows:

(1) Initialization: Set $D_A(B_x, B_y) = d(B_x, B_y)$.
(2) Recursion: For $B_x + 1 \leqslant m_x \leqslant T_x$ and $B_y + 1 \leqslant m_y \leqslant T_y$, compute

$$D_A(m_x, m_y) = \min_{(m'_x, m'_y)} [D_A(m'_x, m'_y) + \varsigma((m'_x, m'_y), (m_x, m_y))], \quad (3)$$

where the incremental distortion $\varsigma((m'_x, m'_y), (m_x, m_y))$ and the intermediate point (m'_x, m'_y) along three local paths $\wp_1$, $\wp_2$, and $\wp_3$ are given in Table 1.
(3) Path backtracking: According to the optimal DTW path, we define the time-scaling factor $\rho(m) = 0.5$, 1, or 2, for the case where the move from the point (m'_x, m'_y) to the point (m_x, m_y) is via the local path $\wp_1$, $\wp_2$, or $\wp_3$, respectively.

Table 1
Incremental distortions and slope weights for local paths



Path	(m'_x, m'_y)	$\varsigma((m'_x, m'_y), (m_x, m_y))$
$\wp_1$	$(m_x - 2, m_y - 1)$	$\frac{1}{2}d(m_x - 1, m_y) + \frac{1}{2}d(m_x, m_y)$
$\wp_2$	$(m_x - 1, m_y - 1)$	$d(m_x, m_y)$
$\wp_3$	$(m_x - 1, m_y - 2)$	$\frac{1}{2}d(m_x, m_y - 1) + \frac{1}{2}d(m_x, m_y)$

4. Pitch modification

The four basic Mandarin tones mentioned earlier have distinctive shapes of F0 contours, whose perception is correlated with the starting frequency, the initial fall and the timing when the turning point appears, as involved in tones 2 and 3 (Shen and Lin, 1991). Our teenage data supported the general conclusion with a different measure. Specifically, instead of focusing on the interactions between the frequency and temporal aspects, we recorded the frequency differences between the highest and the lowest point found on the contours. The results showed a clear trend for the normal speakers with the difference increased when going from tone 1 to tone 4 (e.g., 19.6, 24.8, 53, 113.1 Hz), which was less orderly (e.g., 9.3, 1.6, 28, 66.4 Hz) for the impaired speakers. The most frequent perceptual mistakes made by our impaired speakers were substitutions of tone 3 with tone 2, which left only three perceptual categories 1, 2, and 4 in their tonal repertoire. Unstable tonal productions across recorded tokens were also common.

Most current approaches to voice conversion make little or no use of pitch measures, despite evidence showing that intonational information is highly correlated to speech individuality. The main reason for this is the difficulty in finding an appropriate feature set that captures linguistically relevant intonational information. This problem is alleviated in Mandarin speech conversion task as its tonal system allows relatively non-overlapping characterizations of the corresponding F0 contour dynamics. Speech enhancement can therefore be realized by a proper analysis and control of the F0 contour dynamics. Since pitch is defined only for voiced speech, the pertinent tone-related portions of syllables are the vowel or diphthong nuclei from which distinctive pitch changes are perceived. Recognizing this, we need only to concatenate F0 values of the final subsyllable into a vector and represent it by a small linguistically motivated parameter set. Unlike the conventional frame-based VQ approaches (Abe et al., 1988), this segment-based approach makes it possible to convert not only the static characteristics but also the dynamic characteristics of F0 contours.

Choosing an appropriate representation of F0 contour is the first step in applying pitch modification to the voice conversion. By taking advantage of the simple tone structure of F0 contours in mandarin speech, the polynomial curve fitting technique is used to decompose the F0 contour into mutually orthogonal components in transform domain (Chen and Wang, 1990). The F0 contour can therefore be represented by a smooth curve formed by orthogonal expansion using some low order transform coefficients. In describing the source speaker's F0 contour, F0 are measured only for the final subsyllable and are in the form of $\{w_0(m_x), B_x \leqslant m_x \leqslant T_x\}$. For notational convenience, the F0 contour of a segment with $I_x + 1$ frames is rewritten as $\{w_0(i_x), 0 \leqslant i_x \leqslant I_x\}$, where $i_x = m_x - B_x$ and $I_x = T_x - B_x$. Parameters for pitch modification are then extracted from the F0 contour segment by the orthogonal polynomial transform:

$$b_j^{(x)} = \frac{1}{I_x + 1} \sum_{i_x=0}^{I_x} w_0(i_x) \cdot \Psi_j\left(\frac{i_x}{I_x}\right), \quad j = 0, 1, 2, 3. \tag{4}$$

Due to the smoothness of an F0 contour segment (Chen and Wang, 1990), the first four discrete Legendre polynomials are chosen as the basis functions $\Psi_j(\cdot)$ to represent it. Based on this orthogonal polynomial representation, the source F0 contour is characterized by a 4-dimensional feature vector, $\boldsymbol{b}^{(x)} = (b_0^{(x)}, b_1^{(x)}, b_2^{(x)}, b_3^{(x)})^{\mathrm{T}}$, which will be quantized using vector quantization (VQ) technique. Similarly,

$\boldsymbol{b}^{(y)} = (b_0^{(y)}, b_1^{(y)}, b_2^{(y)}, b_3^{(y)})^{\mathrm{T}}$ is a feature vector representing the F0 contour of the target speaker.

Our conversion technique is based on the codebook mapping and consists of two steps: a learning step and a conversion-synthesis step. In the learning step, the source and target F0 codebooks were separately generated using an orthogonal polynomial representation of F0 contours in training utterances. Each of the two codebooks includes 16 codevectors and is designed using the well-known LBG algorithm (Linde et al., 1980). Next, a histogram of correspondence between codebook elements of the two speakers is calculated. Using this histogram as a weighting function, the mapping codebook is defined as a linear combination of target F0 codevectors. In the conversion-synthesis step, the F0 contour of input speech was orthogonally transformed and vector-quantized using the source F0 codebook. Then, the pitch modification was carried out by decoding them using the mapping codebook. If the decoded codevector is $\hat{\boldsymbol{b}} = (\hat{b}_0, \hat{b}_1, \hat{b}_2, \hat{b}_3)^{\mathrm{T}}$, the modified F0 for frame $m_x = i_x + B_x$ can be approximated as

$$w_0'(i_x + B_x) = \sum_{j=0}^{3} \hat{b}_j \cdot \Psi_j\left(\frac{i_x + B_x}{I_x}\right), \quad 0 \leqslant i_x \leqslant I_x. \tag{5}$$

5. Spectral conversion

In addition to prosodic conversion, the general voice conversion task also necessitates a mapping of spectral envelopes from one speaker to another. Mandarin is a syllable-timed language in which each syllable consists of an initial part and a final part. The primary difficulties in the recognition of Mandarin syllables are tied to the durational differences between the syllable-initial and syllable-final part. Specifically, the initial part of a syllable is short when compared with the final part, which usually causes distinctions among the initial consonants in different syllables to be swamped by the following irrelevant differences among the finals. This may help explain why early approaches that used whole-syllable models as the conversion units did not produce satisfactory results for Mandarin speech conversion. To circumvent this pitfall, we perform spectral conversion only after decomposing the Mandarin syllables into smaller sound units as in phonetic classes.

The acoustic features included in the conversion are cepstral coefficients derived from the smoothed spectrum. The conversion system design involves two essential problems: (1) developing a parametric model representative of the distribution of cepstral coefficients, and (2) mapping the spectral envelopes of the source speaker onto those of the target. In the context of spectral transformation, Gaussian mixture models (GMMs) have been shown to provide superior performance to other approaches based on VQ or neural networks (Stylianou et al., 1998). Our approach began with a training phase in which all cepstral vectors of the same phonetic class were collected and used to train the corresponding GMM associated with the phonetic class by a supervised learning procedure. We consider that the available data consists of two sets of time-aligned cepstral vectors $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$, corresponding, respectively, to the spectral envelopes of the source and the target speakers. The GMM assumes that the probability distribution of the cepstral vectors $\boldsymbol{x}$ takes the following parametric form

$$p(\boldsymbol{x}) = \sum_{i=1}^{I} \alpha_i \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i^{xx}), \tag{6}$$

where α_i denotes a weight of class i, $I = 24$ denotes the total number of Gaussian mixtures, and $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i^{xx})$ denotes the normal distribution with mean vector $\boldsymbol{\mu}_i^x$ and covariance matrix $\boldsymbol{\Sigma}_i^{xx}$. It therefore follows the Bayes theorem that a given vector $\boldsymbol{x}$ is generated from the ith class of the GMM with the probability:

$$h_i(\boldsymbol{x}) = \frac{\alpha_i \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i^{xx})}{\sum_{j=1}^{I} \alpha_j \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_j^x, \boldsymbol{\Sigma}_j^{xx})}. \tag{7}$$

With this, cepstral vectors are converted from the source speaker to the target speaker by the conversion function that utilizes feature parameter correlation between the two speakers. The conversion function that minimizes the mean squared error between converted and target cepstral vectors was given by (Stylianou et al., 1998),

$$\mathcal{F}(\boldsymbol{x}_t) = \sum_{i=1}^{I} h_i(\boldsymbol{x}_t)[\boldsymbol{\mu}_i^y + \boldsymbol{\Sigma}_i^{yx}(\boldsymbol{\Sigma}_i^{xx})^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_i^x)], \tag{8}$$

where for class i, $\boldsymbol{\mu}_i^y$ denotes the mean vector for the target cepstra, $\boldsymbol{\Sigma}_i^{xx}$ denotes covariance matrix for the source cepstra, and $\boldsymbol{\Sigma}_i^{yx}$ denotes the cross-covariance matrix.

Within the GMM framework, training the conversion function can be formulated as one of the optimal estimation of model parameters $\lambda = \{\alpha_i, \boldsymbol{\mu}_i^x, \boldsymbol{\mu}_i^y, \boldsymbol{\Sigma}_i^{xx}, \boldsymbol{\Sigma}_i^{yx}\}$. Our approach to parameter

estimation is based on fitting a GMM to the probability distribution of the joint vector $\boldsymbol{z}_t = [\boldsymbol{x}_t, \boldsymbol{y}_t]^{\mathrm{T}}$ for the source and target cepstra. Covariance matrix $\boldsymbol{\Sigma}_i^z$ and mean vector $\boldsymbol{\mu}_i^z$ of class i for joint vectors can be written as

$$\boldsymbol{\Sigma}_i^z = \begin{bmatrix} \boldsymbol{\Sigma}_i^{xx} & \boldsymbol{\Sigma}_i^{xy} \\ \boldsymbol{\Sigma}_i^{yx} & \boldsymbol{\Sigma}_i^{yy} \end{bmatrix}, \quad \boldsymbol{\mu}_i^z = \begin{bmatrix} \boldsymbol{\mu}_i^x \\ \boldsymbol{\mu}_i^y \end{bmatrix}. \tag{9}$$

The expectation-maximization (EM) algorithm (Dempster et al., 1977) is applied here to estimate the model parameters which guarantees a monotonic increase in the likelihood. Starting with an initial model λ, the new model $\bar{\lambda}$ is estimated by maximizing the auxiliary function

$$Q(\lambda, \bar{\lambda}) = \sum_{t=1}^{T} \sum_{i=1}^{I} p(i|\boldsymbol{z}_t, \lambda) \cdot \log p(i, \boldsymbol{z}_t|\bar{\lambda}), \tag{10}$$

where

$$p(i, \boldsymbol{z}_t|\bar{\lambda}) = \bar{\alpha}_i \mathcal{N}(\boldsymbol{z}_t; \bar{\boldsymbol{\mu}}_i^z, \bar{\boldsymbol{\Sigma}}_i^z), \tag{11}$$

and

$$p(i|\boldsymbol{z}_t, \lambda) = \frac{\alpha_i \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_i^z, \boldsymbol{\Sigma}_i^z)}{\sum_{j=1}^{I} \alpha_j \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_j^z, \boldsymbol{\Sigma}_j^z)}. \tag{12}$$

On each EM iteration, the reestimation formulas derived for individual parameters of class i are of the form

$$\bar{\alpha}_i = \frac{1}{T} \sum_{t=1}^{\mathrm{T}} p(i|\boldsymbol{z}_t, \lambda), \tag{13}$$

$$\bar{\boldsymbol{\mu}}_i^z = \frac{\sum_{t=1}^{\mathrm{T}} p(i|\boldsymbol{z}_t, \lambda)(\boldsymbol{z}_t)}{\sum_{t=1}^{\mathrm{T}} p(i|\boldsymbol{z}_t, \lambda)}, \tag{14}$$

$$\bar{\boldsymbol{\Sigma}}_i^z = \frac{\sum_{t=1}^{\mathrm{T}} p(i|\boldsymbol{z}_t, \lambda)(\boldsymbol{z}_t - \boldsymbol{\mu}_i^z)(\boldsymbol{z}_t - \boldsymbol{\mu}_i^z)^{\mathrm{T}}}{\sum_{t=1}^{\mathrm{T}} p(i|\boldsymbol{z}_t, \lambda)}. \tag{15}$$

The new model $\bar{\lambda}$ then becomes λ for the next iteration and the reestimation process is repeated until the likelihood reaches a fixed value.

6. Experimental results

Experiments were carried out to investigate the potential advantages of using the proposed conversion algorithms to enhance the hearing-impaired Mandarin speech. Our efforts began with the collection of a speech corpus that contained two sets of monosyllabic utterances, one for system learning and one for testing in our voice conversion experiment. The text material consisted of 76 isolated tonal CV syllables (19 base syllables × 4 tones), formed by pairing the three prominent vowels /a,i,u/ with 11 consonants, the five fricatives and the six affricates of Mandarin Chinese, but excluding combinations that were phonologically unacceptable. The choice of these two classes was based on the research findings showing these consonants appeared as the most frequently misarticulated sounds made by the hearing-impaired Mandarin speakers (Lee, 1999). Speech samples were produced by two male adult speakers, one with normal hearing sensitivity and the other with congenital severe-to-profound (>70 dB) hearing loss. The speech of the impaired speaker was largely intelligible in sentences but often caused misunderstanding if produced in syllable forms due to prosodic deviations and misarticulated initial consonants.

Fig. 3 presents the results of our pitch modification method for transforming F0 contours. Panels 3(a) and 3(c) are the F0 contours for the source and the target syllable /ti/ spoken with four different tones, and panel 3(b) is the converted F0 contour using VQ and orthogonal polynomial representation. Comparison of F0 variations as a function of time found in panel 3(b) with 3(a) clearly shows the improvements on tones 2 and 3. Our next examination focused on how the converted F0 contours were perceived in relation to those of the source. For easy judgments of the tonal categories, only syllables with one consonant class (affricate) were used, with a total of 40 tonal syllables (10 for each tone). Four male and one female adult native speakers of Mandarin Chinese, all with normal hearing status, served as the listeners. Tables 2 and 3 present the confusion matrices showing the tone recognition results for the source and the converted set, respectively. The results in each table were based on the listeners' judgments of 400 responses (40 tonal syllables × 5 listeners × 2 sessions). It is clear that the proposed system resulted in more intelligible stimuli with an average tone recognition score of 86.25%, compared with 69% for the source stimuli. The results further showed an improvement of 38% and 28% for syllables with tone 2 and tone 3, respectively.

To establish the statistical significance of these results, we calculated the P-value using a Z-test (Johnson and Bhattacharyya, 1996). If we let p_1 and p_2 denote the recognition rates for the source and converted set, respectively, our objective was to test the null hypothesis $H_0 : p_1 \geqslant p_2$. Based on the statistics in Table 4, the Z-test yielded a small

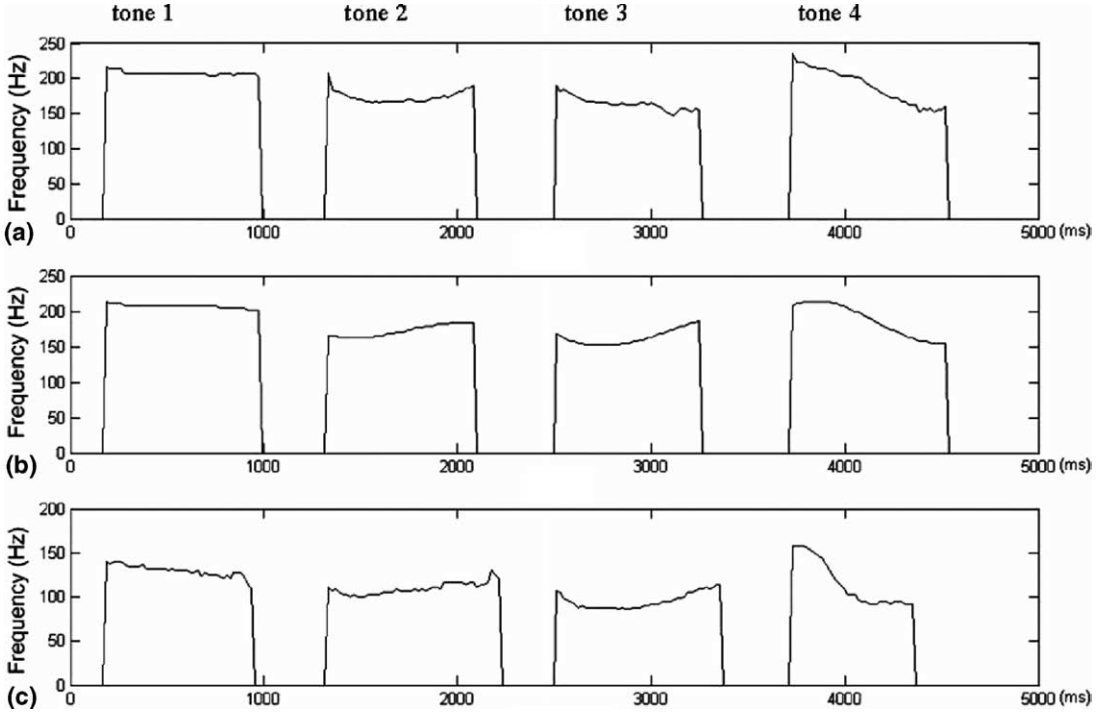


Fig. 3. F0 contours for syllable /ti/ spoken with four different tones: (a) source speech, (b) converted speech, and (c) target speech.

Table 2
Confusion matrix showing tone recognition results for source syllables

Response	Stimulus			
	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	[98]	32	16	0
Tone 2	0	[43]	35	1
Tone 3	2	23	[45]	9
Tone 4	0	2	4	[90]

Table 3
Confusion matrix showing tone recognition results for converted syllables

Response	Stimulus			
	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	[97]	0	1	0
Tone 2	3	[81]	26	1
Tone 3	0	19	[73]	5
Tone 4	0	0	0	[94]

P-value ($P < 0.0002$); therefore, the null hypothesis was strongly rejected. Further evidence of improvement is seen on Fig. 4, which shows our prosodic modification applied to continuous speech. A four-syllable utterance, containing tones 4-4-3-3, was used. According to the tone-sandhi rule, the first tone 3 should be produced with a tone 2 F0 pattern. The audio presentation, however, showed that the first tone 3 was produced more like tone 1 than the targeted tone 2. A comparison of the F0 contours for the source and the target utterances showed that the former exhibited fewer fine fluctuation details, even though the variation ranges were both within 100 Hz. Further, the first tone 4 was essentially carrying a tone 1 F0 pattern and the last tone 3 was produced with the rising part truncated. The improvement due to prosodic modification can be seen in the following areas. First, the missing falling part in the fist tone 4 and the dipping of the last tone 3 were fully restored. Second, the rising part of

Table 4
Raw data and tone recognition rates derived from Tables 2 and 3

	Number of correct identification	Number of wrong identification	Recognition rate
Source stimuli	276	124	$p_1 = \frac{276}{400}$
Converted stimuli	345	55	$p_2 = \frac{345}{400}$

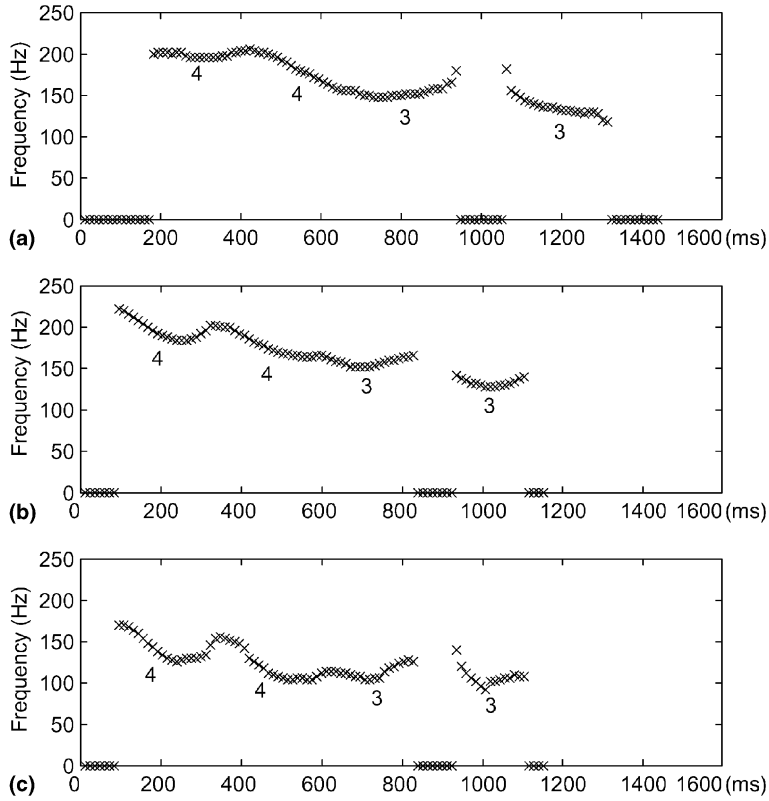


Fig. 4. F0 contours for a four-syllable phrase /ying-4 yong-4 ruan-3 ti-3/ (meaning *application software*): (a) source speech, (b) converted speech, and (c) target speech.

the first tone 3 segment was steeper in slope, making it more appropriate for the targeted tone 2. To hear audio examples of the voice conversion system, please visit the web site at http://a61.cm.nctu.edu.tw/demo.

Results of the spectral conversion were analyzed acoustically with software spectrograph to assess how closely the converted speech resembled the target speech in rendering acoustic cues for phoneme perception. The improvement for the fricatives is shown in three aspects: (1) lengthening of the consonant duration, (2) a less abrupt transition, or a gradual blending of the acoustic energy, near the consonant-vowel boundary, and (3) a redistribution of acoustic energy around appropriate frequency regions, such as an elevation to 3 kHz for the syllable /shu/ or to 4 kHz for the syllable /shii/. An example of such spectral differences for the syllable /shu/ is shown in Fig. 5. Even closer spectrographic matches were obtained for the affricates, as shown in Fig. 6 using /chii/ as an example. In normal production, affricates are stops followed by fricatives, which are individually represented on the spectrograph as a burst with its energy concentrated at higher frequencies to be blended immediately with those of the following fricative. The distorted affricate, however, was translated spectrographically into a stop that included a full voicing gap but not much of frication. Our analysis revealed that the conversion filled the gap, softened the burst, removed the low frequency energy and elevated the fricative portion to normal frequency ranges. When examined along with audio presentations, this modification also resulted in a change of the vowel percept from the erroneous, high front but lip-rounding, vowel /yu/ to the correct /i/, even though formant modification for the vowel was less apparent.

Two listening tests, preference and intelligibility, were conducted to determine whether the above spectrographic enhancement could also be realized perceptually. The five listeners for the previous tone recognition test were used. In the preference test, the listeners were asked to give their preference judgments over pairs of source vs. converted syllables. A two-alternative-forced-choice (2AFC) test paradigm was used, in which the presentation order of

the two stimuli was randomized. For converted stimuli, two sets of converted syllables were used: (1) those with spectral conversion only and (2) those with combined spectral and time-scaled conversions. The results showed 62% of the 380 responses (2 stimulus sets × 19 base syllables × 5 listeners × 2 sessions) preferred spectrally modified syllables to source syllables, while 84% preferred those with combined modifications. To further validate the effect of the proposed approach, intelligibility

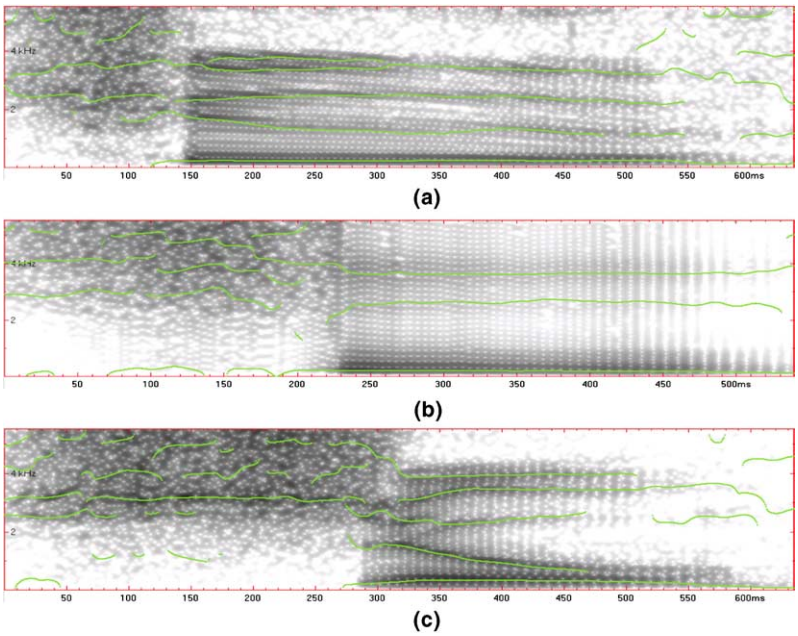


Fig. 5. Spectrograms for syllable /shu/: (a) source speech, (b) converted speech, and (c) target speech.

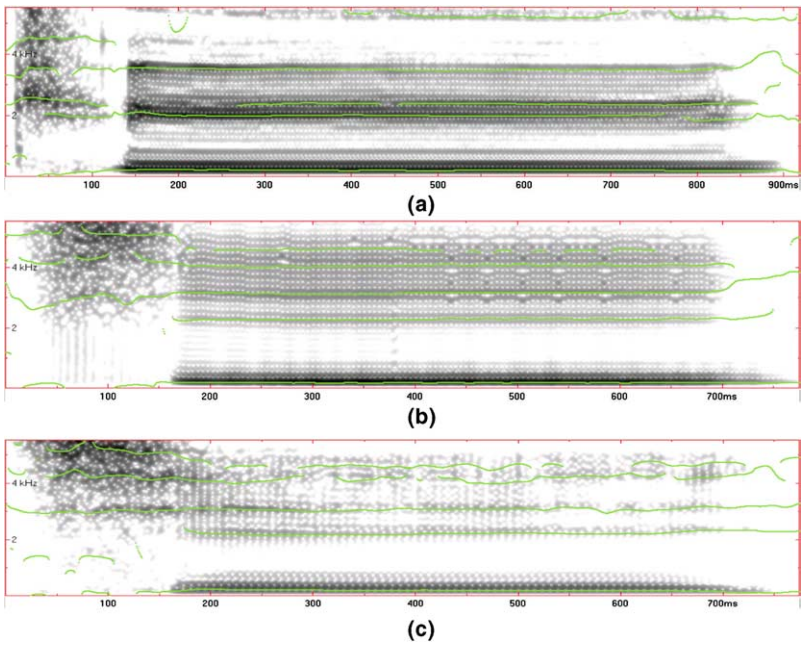


Fig. 6. Spectrograms for syllable /chii/: (a) source speech, (b) converted speech, and (c) target speech.

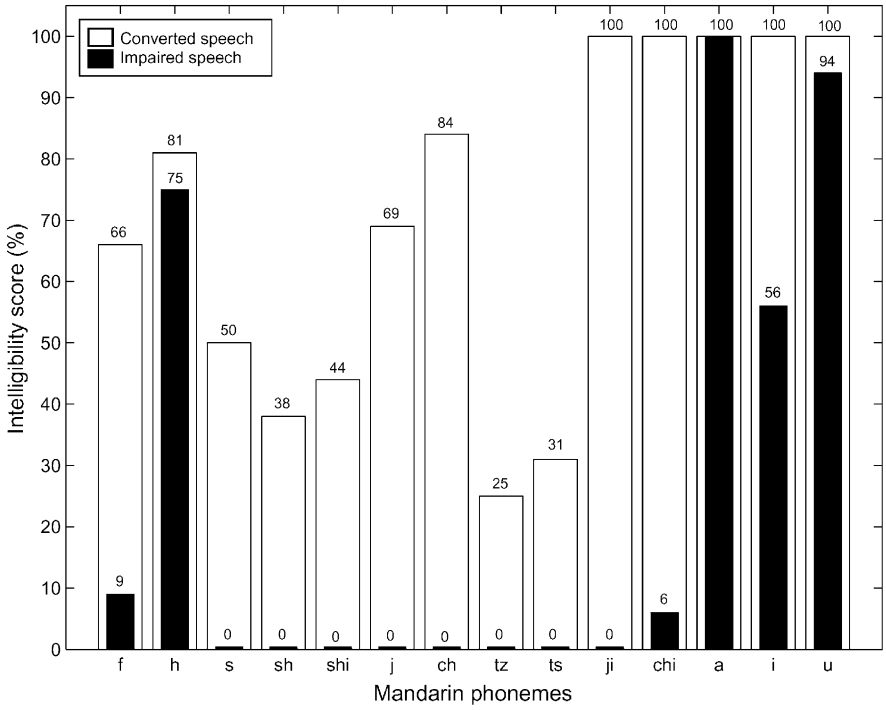


Fig. 7. Percent correct phoneme recognition scores for source and converted speech.

measures were obtained for 19 base syllables before and after spectral conversion. The listeners were instructed to write down their responses using Mandarin phonetic symbols. Fig. 7 shows comparison of the percent correct phoneme recognition scores for the source and the converted stimuli. Individual phonemes were arranged from left to right into three groups, fricative, affricate, and vowel. Recognition of vowels /a,u/ was near perfect even without the modification. In contrast, recognition for the affricates and the fricatives (with the exception of /h/) was either near or at 0%, a finding consistent with our earlier observation that these two consonant classes are frequently substituted with stops by the hearing-impaired speakers. The relatively good recognition for /h/, even for the source, could be explained by the fact that little oral modification of the glottal air source was required during articulation. With the converted stimuli, an improvement was seen in all three groups. An average increase of 47.25% was obtained for the fricatives, with /h/ counted out. The amount was further increased by 20% (=67.17%) for the affricates, with /ji, chi/ showing a total correction, making this group the phoneme class that benefited the most from our application. The vowels, despite their small improvement, were the only group showing a total correction for all its members.

7. Conclusions

This study presents a novel means of exploiting spectral and prosodic transformations in enhancing disordered speech. In spectral conversion, subsyllable-based GMMs were applied within the sinusoidal framework to modify the articulation-related parameters of speech. In prosodic conversion, we found the tone structure of F0 contour in Mandarin speech could be used to advantage in orthogonal polynomial representation of pitch contours. The results also suggest a new approach to time-scaling modification in which the initial part of a syllable is linearly normalized with a fixed factor, and then a DTW algorithm is used to control the time-varying scaling factor for the final part. Evaluations by objective tests and listening tests show that the proposed techniques can improve the intelligibility and naturalness of the hearing-impaired Mandarin speech. Although fairly good performances were reported in these experiments, more work is needed to further validate the proposed voice conversion system for a wider range of hearing-impaired speech

corpora. For example, in extending the current system to continuous speech, more sophisticated tone models may be needed as tone patterns of syllables in continuous speech are subject to various modifications by sandhi rules.

Acknowledgement

This study was supported by the National Science Council, Taiwan, Republic of China, Under Contracts NSC 93-2213-E-009-123 and NSC 93-2614-H-134-001-F20.

References

Abe, M., Nakamura, S., Shikano, K., Kuwabara, H., 1988. Voice conversion through vector quantization. In: Proc. ICASSP'88, pp. 655–658.

Bi, N., Qi, Y., 1997. Application of speech conversion to alaryngeal speech enhancement. IEEE Trans. Speech Audio Process. 5, 97–105.

Chang, B.L., 2000. The perceptual analysis of speech intelligibility of students with hearing impairments. Bull. Special Education 18, 53–78.

Chen, S.H., Wang, Y.R., 1990. Vector quantization of pitch information in Mandarin speech. IEEE Trans. Communications 38, 1317–1320.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. 39, 1–38.

Hochberg, I., Levitt, H., Osberger, M.J., 1983. Speech of The Hearing Impaired: Research, Training, and Personnel Preparation. University Park Press, Maryland.

Johnson, R.A., Bhattacharyya, G.K., 1996. Statistics: Principles and Methods. John Wiley and Sons, New York.

Kain, A., Macon, M.W., 1998. Spectral voice conversion for text-to-speech synthesis. In: Proc. ICASSP'98, pp. 285–288.

Lee, L.S., 1997. Voice dictation of Mandarin Chinese. IEEE Signal Process. Mag., 63–101.

Lee, P.C., 1999. A study on acoustic characteristic of Mandarin affricates of hearing-impaired speech. Bull. Special Eduation Rehabil. 7, 79–112.

Lin, B.G., Huang, Y.C., 1997. An analysis on the hearing impaired students' Chinese language abilities and its error patterns. Bull. Special Educ. 15, 109–129.

Linde, Y., Buzo, A., Gray, R.M., 1980. An algorithm for vector quantizer design. IEEE Trans. Communications 28, 84–95.

McAulay, R.J., Quatieri, T.F., 1995. Sinusoidal Coding: Speech coding and synthesis. Elsevier, Amsterdam.

McGarr, N.S., Harris, K.S., 1983. Articulatory control in deaf speaker. In: Hochberg, I., Levitt, H., Osberger, M.J. (Eds.), Speech of the Hearing Impaired. University Park Press, Baltimore.

Monsen, R., 1978. Toward measuring how well hearing-impaired children speak. J. Speech Hearing Res. 21, 197–219.

Ohde, R.N., Sharf, D.J., 1992. Phonetic Analysis of Normal and Abnormal Speech. Merrill, New York.

Oppenhein, A.V., Schafer, R.W., 1989. Discrete-time Signal Processing. Prentice Hall, New Jersey.

Osberger, M.J., Levitt, H., 1979. The effect of timing errors on the intelligibility of deaf children's speech. J. Acoust. Soc. Amer. 66 (5), 1316–1324.

Quatieri, T.F., McAulay, R.J., 1992. Shape invariant time-scale and pitch modification of speech. IEEE Trans. Signal Process. 40 (3), 497–510.

Rabiner, L., Juang, B.H., 1993. Fundamentals of Speech Recognition. Prentice Hall, New Jersey.

Shen, X.S., Lin, M., 1991. A perceptual study of Mandarin tones 2 and 3. Lang. Speech 34 (2), 145–156.

Stylianou, Y., Cappe, O., Moulines, E., 1998. Continuous probabilistic transform for voice conversion. IEEE Trans. Speech Audio Process. 6, 131–142.