Application of Voice Conversion to Hearing-Impaired Mandarin Speech Enhancement

Chen-Long Lee[*]*, Wen-Whei Chang*[*]*, and Yuan-Chuan Chiang*[†]

[*]Department of Communications Engineering
National Chiao-Tung University
Hsinchu, Taiwan, Republic of China

[†]Department of Special Education
National Hsinchu Teachers College
Hsinchu, Taiwan, Republic of China

Abstract

This paper studies the application of voice conversion to hearing-impaired Mandarin speech enhancement. The system is based on the combined use of a sinusoidal analysis-synthesis model and *a priori* knowledge about Mandarin syllable phonetic structures. We propose a time-scale modification algorithm that finds accurate alignments between hearing-impaired and normal utterances. Using the alignments, spectral conversion is performed by a continuous probabilistic transform based on a Gaussian mixture model. Simulation results indicate that the proposed system can improve the intelligibility of hearing-impaired Mandarin speech.

1 Introduction

The speech of hearing-impaired speakers suffers from misarticulations and prosodic deviations [1,2], which reduces their intelligibility and restricts their use of any voice-controlled electronic devices. This motivates our research into trying to devise a voice converter that modifies the speech of a hearing-impaired (source) speaker to be perceived as if it was uttered by a normal (target) speaker. The key to voice conversion lies in the detection and exploitation of characteristic features that distinguish the impaired speech from the normal speech at phonetic and prosodic levels [3]. Phonetic features are encoded in the spectral envelope, whereas prosodic information can be found in pitch and duration variations that span across segments.

The characteristics of Mandarin, significantly different from those of alphabetic western languages, lead to the fact that conversion techniques that consider Chinese language characteristics are believed to be the key to providing better solutions to the problem. Mandarin syllables are traditionally decomposed into *initials* and *finals*, in which *initial* means the consonant onset of a syllable while *final* means the vowel or diphthong, but including an optional medial or nasal ending. The primary difficulties in Mandarin pronunciation are caused by the existence of 38 confusing sets, each of which consists of syllables sharing the same *final* but with different *initials*. It was found [4] that the hearing-impaired speakers made twice as many consonant as vowel errors, and further, that the most common errors in the consonants were affricates and fricatives. The hearing-impaired speech also contains numerous timing errors including a reduced speaking rate, excessive shortening of consonants, and insertion of long pauses. Thus there is a need to apply spectral conversion as well as time-scale modification in order to achieve hearing-impaired Mandarin speech enhancement.

2 System Implementation

The voice conversion system has four major components: speech analysis, time-scale modification, spectral conversion, and speech synthesis. A block diagram of the proposed system is shown in Fig. 1. Speech analysis is based on a harmonic sine-wave model [5] that decomposes speech signals into the product of excitation and system spectra, and then represents the excitation signal by a sum of sine waves whose frequencies are integer multiples of the pitch frequency. Sinusoidal representation of speech is performed frame by frame and is of the following form:

$$s(n) = \sum_{k=1}^{L(m)} a_k(m) M_k(m) cos[\Omega_k(m) + \Phi_k(m)] \qquad (1)$$

where for the mth frame, $L(m)$ is the number of sine waves, $a_k(m)$ and $\Omega_k(m)$ represent the excitation amplitude and phase, $M_k(m)$ and $\Phi_k(m)$ represent the system amplitude and phase, respectively.

A two-step procedure is used to compute the excitation phase $\Omega_k(m)$ of the kth sine wave. First, the pitch periods $P(m)$ are accumulated until a pitch pulse crosses the center of the mth frame of duration $Q(m)$. The location of this pulse is the onset time at which sine waves are in phase and can be written as

$$n_0(m) = n_0(m-1) + J_m P(m), \qquad (2)$$

where J_m corresponds to the pulse closest to the center of the mth frame. Then, the excitation phase is given by

$$\Omega_k(m) = -[mQ(m) - n_0(m)] w_k(m), \qquad (3)$$

where $w_k(m) = kw_0(m)$ is the frequency of the kth sine-wave and $w_0(m) = 2\pi/P(m)$ is the pitch frequency. The vocal tract transfer function can be described in terms of its amplitude envelope $\hat{M}(w;m)$ and phase envelope $\hat{\Phi}(w;m)$. The system amplitude $M_k(m)$ and phase $\Phi_k(m)$ are then given by samples of their respective envelopes at the frequency $w_k(m)$, i.e., $M_k(m) = \hat{M}(w_k;m)$ and $\Phi_k(m) = \hat{\Phi}(w_k;m)$.

Following the speech analysis, sine-wave amplitudes $A_k(m) = a_k(m)M_k(m)$ were used to compute 25-dimensional cepstral vectors for spectral conversion. The *initial-final* boundary of a Mandarin syllable was determined by the voicing probability P_v that is a measure of how well the harmonic set of sine waves fits the measured set of sine waves and was determined as part of the pitch estimation process. Using subsyllables as the basic units, the voice conversion involves the manipulations of functions which describe the time evolution of the excitation and system contributions of the amplitude and phase of each sine-wave component. In the synthesis procedure, the modified excitation and system amplitudes are multiplied and linearly interpolated over consecutive frames. Also, the modified excitation and system phases are summed and interpolated via the cubic phase interpolator [3].

3 Time-Scale Modification

Hearing-impaired speech is generally characterized by a much lower speaking rate and by excessive shortening of consonants. Thus there is a need to normalize out speaking rate variation as well as duration variation in order for the frame correspondence to be meaningful before spectral conversion can be made. The first step consists in scaling the synthesis frame duration by a factor $\rho(m)$, that is $Q'(m) = \rho(m)Q(m)$. The case $\rho > 1$ corresponds to time-scale expansion, and $\rho < 1$ corresponds to time-scale compression. The onset time $n_0'(m)$ is then obtained relative to the center of the new synthesis frame of duration $Q'(m)$. The change in onset time also corresponds to modification of the excitation phase $\Omega_k'(m)$ of each underlying sine-wave as follows:

$$\Omega_k'(m) = -[mQ'(m) - n_0'(m)]w_k(m) \tag{4}$$

We consider two sets of paired cepstral vectors $\mathbf{x}(i_x)$ and $\mathbf{y}(i_y)$ corresponding, respectively, to the same syllable uttered by the source and the target speakers. Cepstral features of the source speaker are denoted by $\mathbf{x}_1^{T_x} = \{\mathbf{x}(i_x), i_x = 1, 2, \ldots, T_x\}$, where T_x is the duration in frames. Similarly, $\mathbf{y}_1^{T_y} = \{\mathbf{y}(i_y), i_y = 1, 2, \ldots, T_y\}$ is the sequence of T_y vectors representing the cepstral features of the target speaker. Let B_x and B_y represent the starting frame for the *final* subsyllable in the source and target utterances, respectively. The local distortion between vectors $\mathbf{x}(i_x)$ and $\mathbf{y}(i_y)$ is defined by a squared Euclidean distance, i.e., $d(\mathbf{x}(i_x), \mathbf{y}(i_y)) = \|\mathbf{x}(i_x) - \mathbf{y}(i_y)\|^2$.

Different normalization approaches were applied in the time-intervals where the frames corresponding to both speakers were marked as *initial* or *final* subsyllables. For the *initial* subsyllables, a linear time normalization was applied to $\mathbf{x}_1^{B_x-1}$ with a fixed rate change $\rho = (B_y - 1)/(B_x - 1)$. For the *final* subsyllables, the computed cepstral vectors were time aligned between the source and target speakers using the procedure of dynamic time warping (DTW) [6]. The DTW alignment between paired patterns $\mathbf{x}_{B_x}^{T_x}$ and $\mathbf{y}_{B_y}^{T_y}$ can be formulated as a path finding problem over a set of grid points (i_x, i_y). We first consider a pattern dissimilarity measure $D_A(i_x, i_y)$, representing the minimum partial accumulated distortion along a path connecting (B_x, B_y) and (i_x, i_y). Then, the best path from (B_x, B_y) to (T_x, T_y) is found by the following recursion formula: for $B_x \le i_x \le T_x$ and $B_y \le i_y \le T_y$,

$$D_A(i_x, i_y) = \min_{(i_x', i_y')} [D_A(i_x', i_y') + \varsigma((i_x', i_y'), (i_x, i_y))]. \tag{5}$$

where the intermediate point (i_x', i_y') and incremental distortion $\varsigma((i_x', i_y'), (i_x, i_y))$ along three paths $\wp_1, \wp_2$, and $\wp_3$ are given in Fig. 2. The time-varying rate change has the value ρ=0.5, 1, or 2, for the case where the best path connecting (i_x', i_y') and (i_x, i_y) is via the path $\wp_1, \wp_2$, or $\wp_3$, respectively.

4 Spectral Conversion

Spectral conversion is a feasible technique for modifying articulation-related parameters of speech. Depending on the manner of articulation, phonemes can be categorized into five phonetic classes including affricate, fricative, nasal, stop, and vowel. The spectral conversion system consists of two steps: a learning step and a conversion-synthesis step. In the learning step, phonemes belonging to the same phonetic class were grouped together and characterized under the form of a Gaussian mixture model (GMM). In the conversion-synthesis step, cepstral features of each phoneme were converted using a mapping function that minimized the spectral distortion between the converted speech and the target speech. Suppose that source and target vectors drawn from the same syllable were time-aligned and collected, respectively, into the cepstral sequences $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]$ and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T]$.

In the GMM algorithm, the probability distribution of cepstral vector $\mathbf{x}$ is in the form of

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

where α_i denotes the mixture weight of ith Gaussian component , $\mathcal{N}(\cdot)$ represents a Gaussian density with mean vector $\boldsymbol{\mu}_i^x$ and covariance matrix $\boldsymbol{\Sigma}_i^{xx}$. From this it can be shown that $\mathbf{x}$ is generated from the ith Gaussian component with the probability:

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

The mapping function minimizing the mean square error between the $\mathcal{F}(\mathbf{x}_t)$ and $\mathbf{y}_t$ was given by [7],

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

where for the ith Gaussian component, $\boldsymbol{\mu}_i^y$ denotes the mean vector for target utterances, $\boldsymbol{\Sigma}_i^{xx}$ denotes the covariance matrix for source utterances, and $\boldsymbol{\Sigma}_i^{yx}$ denotes the cross-covariance matrix. The expectation-maximization (EM) algorithm [8] is employed here to estimate model parameters $\lambda = \{\alpha_i, \boldsymbol{\mu}_i^x, \boldsymbol{\mu}_i^y, \boldsymbol{\Sigma}_i^{xx}, \boldsymbol{\Sigma}_i^{yx}\}$.

5 Experimental Results

Experiments were carried out to investigate the potential advantages of using subsyllable-level conversion algorithms to enhance the hearing-impaired Mandarin speech. Our effort 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 materials consisted of 19 monosyllables containing the *final* vowel /i/, /u/ and /a/, and the *initial* consonant was affricate or fricative. Speech samples were produced by two male speakers, one is normal-listening and the other has congenital severe-to-profound ($>$ 70 dB) hearing loss. The hearing-impaired utterances were largely intelligible in sentences but often caused misunderstanding in syllables due to misarticulation of consonant phonemes and improper control of duration. In Fig. 3, a comparison of the duration statistics of Mandarin phonemes for the two speakers in our database are given.

A paired comparison approach was used to determine whether converted utterances sounded more pleasant to the listeners than those uttered by the hearing-impaired. Four native speakers of Mandarin provided the preference judgments. Overall, 62% of the responses prefer utterances converted using only the spectral conversion over impaired utterances. The investigation further showed that a preference score of 84 % was obtained for utterances converted using joint spectral and time-scale modification. To elaborate further, results of the conversion were analyzed with software spectrograph to assess how closely the converted speech resembled the normal speech in rendering acoustic cues for phoneme perception. The best results were seen in the affricate consonants. In normal production, affricates are stops followed by fricatives, while spectrographically the burst of stops only appears shortly and occupies frequencies where the energy of the following fricatives concentrates. The hearing-impaired speech, however, showed affricates that contained a complete stop. Our analyses revealed that the conversion softened the burst, removed their low frequency energy and elevated the fricative portion to normal frequency ranges. An example of such spectral differences for the syllable /chi/ is shown in Fig. 4. In it we also see that vowel /i/ of the converted speech was greatly enhanced by restructuring its formant frequencies to restore formant transition cues as in normal production.

6 Conclusions

This study presents a novel means of exploiting joint spectral and time-scale modification to enhance the hearing-impaired Mandarin speech. By taking advantage of the syllable phonetic structures of Mandarin, durations of *initial* and *final* subsyllables were separately normalized to compensate for the rate of articulation. A GMM-based spectral conversion algorithm was also applied to modify the articulation-related parameters of speech. Evaluation by objective tests and listening tests shows that the proposed techniques can improve the intelligibility of the hearing-impaired Mandarin speech.

Acknowledgements

This study was supported by the National Science Council, Republic of China, under contract NSC 92-2218-E-009-009.

References

[1] Monsen, R., "Toward measuring how well hearing-impaired children speak", *Journal of Speech and Hearing Research*, 21:197-219, 1978.

[2] Massen, B. and Provel, D., "The effect of segmental and suprasegmental corrections on the intelligibility of deaf speech", *J. Acoust. Soc. Am.*, 78:877-886, 1985.

[3] Quatieri T. F. and McAulay R. J., "Shape invariant time-scale and pitch modification of speech", *IEEE Trans. Signal Processing*, vol. 40, pp.497-510, no. 3, Mar. 1992.

[4] Chang, B. L., "The perceptual analysis of speech intelligibility of students with hearing impairments",*International Congress on Education for the Deaf*, Sydney, Australia, July 2000.

[5] McAulay, R. J. and Quatieri, T. F., "Speech analysis-synthesis based on a sinusoidal representation", *IEEE Trans. Acoust., Speech Sig. Process.*, vol. 34:744-754, 1986.

[6] Rabiner, L. and Juang, B. H., *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[7] Stylianou, Y., Cappe, O., and Moulines, E., "Continuous probabilistic transform for voice conversion", *IEEE Trans. Speech and Audio Processing*, vol. 6:131-142, March 1998.

[8] Dempster, A., Laird, N., and Rubin, D., "Maximum likelihood from incomplete data via the EM algorithm", *J. Royal Stat. Soc.*, vol. 39:1-38, 1977.

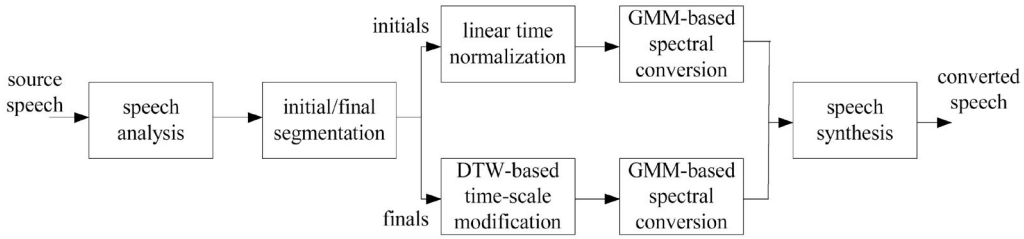
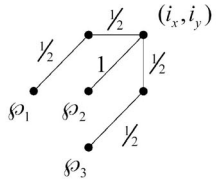


Fig. 1: The voice conversion system.



path	(i_x', i_y')	$\varsigma((i_x', i_y'), (i_x, i_y))$
$\wp_1$	$(i_x - 2, i_y - 1)$	$\frac{1}{2}d(i_x - 1, i_y) + \frac{1}{2}d(i_x, i_y)$
$\wp_2$	$(i_x - 1, i_y - 1)$	$d(i_x, i_y)$
$\wp_3$	$(i_x - 1, i_y - 2)$	$\frac{1}{2}d(i_x, i_y - 1) + \frac{1}{2}d(i_x, i_y)$

Fig. 2: Incremental distortions for paths with local continuity constraints.

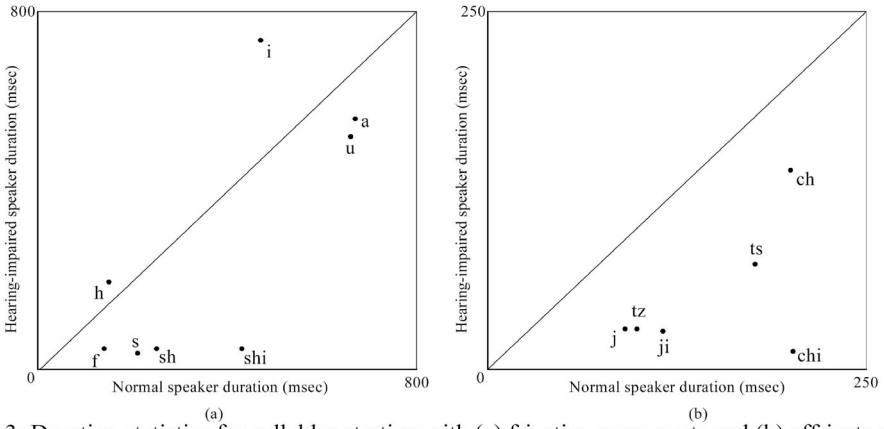
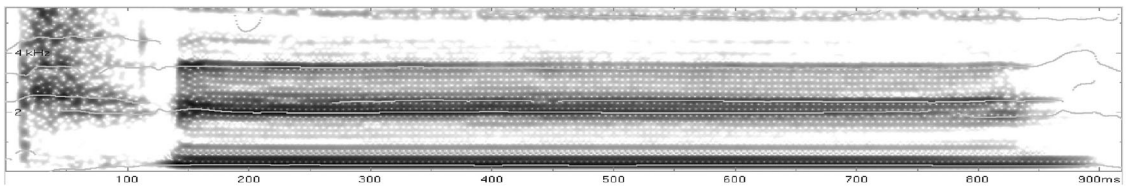
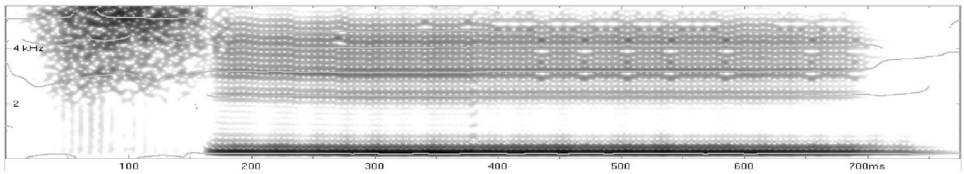


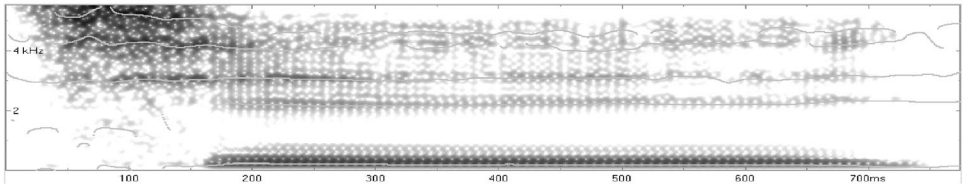
Fig. 3: Duration statistics for syllables starting with (a) fricative consonants and (b) affricate consonants



(a)



(b)



(c)

Fig. 4: Spectrogram comparisons of syllable /chi/. (a) impaired speech, (b) converted speech, (c) normal speech.