Enhancement of Hearing-Impaired Mandarin Speech

Chen-Long Lee, Ya-Ru Yang*, Wen-Whei Chang*, and Yuan-Chuan Chiang*†

* Department of Communication Engineering
National Chiao Tung University
Hsinchu, Taiwan, Republic of China
† Department of Special Education
National Hsinchu Teachers College
Hsinchu, Taiwan, Republic of China
wwchang@cc.nctu.edu.tw

Abstract

This paper presents a new voice conversion system that modifies misarticulations and prosodic deviations of the hearing-impaired Mandarin speech. The basic strategy is the detection and exploitation of characteristic features that distinguish the impaired speech from the normal speech at segmental and prosodic levels. For spectral conversion, cepstral coefficients were characterized under the form of a Gaussian mixture model with parameters converted using a mapping function that minimizes the spectral distortion between the impaired and normal speech. We also proposed a VQ-based approach to prosodic conversion that involves modifying the features extracted from the pitch contour by orthogonal polynomial transform. Experimental results indicate that the proposed system appears useful in enhancing the 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 impaired speech to be perceived as if it was uttered by a normal 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 segmental and prosodic levels. Segmental features that contribute to speech individuality are encoded in the spectral envelope, whereas prosodic information can be found in pitch, energy, and duration variations that span across segments.

Most current approaches to voice conversion [3,4] make little or no use of prosodic features, despite evidence showing that prosodic information is closely related to listeners' perceptions of speech quality. The main reason for this is the difficulty of finding an appropriate feature set that captures linguistically relevant prosodic information. This problem can be alleviated in the Mandarin speech conversion task, mainly because the Chinese tonal system provides an efficient way to describe pitch contour dynamics. Mandarin speech is a tonal language, which implies syllables may have the same phonetic compositions, but different lexical meanings when spoken with different tones. In this study, we propose that speech waveforms are modeled by a sum of sine-waves whose frequencies, amplitudes, and phases are chosen to make the reconstruction a best fit to the original speech [5]. Next, spectral and prosodic conversion techniques were applied on sine-wave parameters to enhance the hearing-impaired Mandarin speech.

2. Sinusoidal Speech Model

We applied the voice conversion on speech signals analyzed by the harmonic sine-wave model, in which frequencies are integer multiples of pitch frequency and amplitudes and phases are chosen to be harmonic samples of the magnitude and phase spectra [5]. Specifically, the general form of a sinusoidal speech model can be expressed as

$$\hat{s}(n) = \sum_{l=1}^{L} A_l cos(nlw_0 + \theta_l) \quad (1)$$

where L denotes the number of sinusoids, w_0 represents the pitch frequency, A_l and θ_l are the amplitude and phase of the l-th sinusoidal component, respectively.

A more efficient representation is achievable by fitting a set of cepstral coefficients to an envelope of the measured sine-wave amplitudes. In this study, the cepstrum order is $q = 25$ using an analysis frame length of 13.6 ms and a sampling rate of 11 kHz. For the system with transfer function $H(z)$, the real cepstrum is defined as the sequence of coefficients in the power series representation of its log magnitude $log|H(w)|$. The main attraction of cepstral representation is that it exploits the minimum-phase model, where the log magnitude $log|H(w)|$ and phase $\Phi(w)$ of the system function can be uniquely related in terms of the Hilbert transform [5]. With this exploitation, additional economies in representing sine-wave phases can be obtained by explicitly identifying the phase components due to the excitation and the vocal system.

3. GMM-Based Spectral Conversion

In the context of speaking aids, a voice converter aims to modify the speech of a hearing-impaired (source) speaker to be perceived as if a normal (target) speaker had spoken it. Figure 1 shows the experimental arrangement of the proposed voice conversion system. Using the sinusoidal analyzer as a front end, sine-wave amplitudes of the speech were used to compute the cepstrum of its smoothed spectrum which is then converted by the mapping function. The acoustic space of a speaker is modeled by a Gaussian mixture model (GMM) and the mapping function is designed to exploit feature parameter correlation between the source speaker and the target speaker [4]. The first

step consists in normalizing different speaking rates in order for the spectral representations to be meaningful before a conversion can be made. Source and target vectors drawn from the same text were time-aligned using a dynamic time warping (DTW) algorithm and collected, respectively, into the 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 a q-dimensional cepstral vector $\mathbf{x}$ is in the form of

$$p(\mathbf{x}) = \sum_{i=1}^{I} \alpha_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{2}$$

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

$$h_i(\mathbf{x}) = \frac{\alpha_i \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i^{xx})}{\sum_{j=1}^{I} \alpha_j \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_j^x, \boldsymbol{\Sigma}_j^{xx})}. \tag{3}$$

The mapping function minimizing the mean square error between the $\mathcal{F}(\mathbf{x}_t)$ and $\mathbf{y}_t$ is given by [4],

$$\mathcal{F}(\mathbf{x}_t) = \sum_{i=1}^{I} h_i(\mathbf{x}_t)[\boldsymbol{\mu}_i^y + \boldsymbol{\Sigma}_i^{yx}(\boldsymbol{\Sigma}_i^{xx})^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_i^x)], \tag{4}$$

where $\boldsymbol{\mu}_i^x$ and $\boldsymbol{\mu}_i^y$ denote mean vectors of class i for the source and target speakers, $\boldsymbol{\Sigma}_i^{xx}$ denotes covariance matrix of class i for the source speaker, and $\boldsymbol{\Sigma}_i^{yx}$ denotes the cross-covariance matrix of class i for the source and target speakers. The expectation-maximization (EM) algorithm [6] is employed here to estimate model parameters $\lambda = \{\alpha_i, \boldsymbol{\mu}_i^x, \boldsymbol{\mu}_i^y, \boldsymbol{\Sigma}_i^{xx}, \boldsymbol{\Sigma}_i^{yx}\}$.

4. VQ-Based Prosodic Conversion

Early approaches that made little or no use of prosodic features did not produce satisfactory results in the Mandarin speech conversion task. Further enhancement can be realized by better modeling of prosody and by additionally incorporating prosodic conversion into the voice conversion system. There are five basic tones in Mandarin speech, namely, Tone 1 (high-level), Tone 2 (midrising), Tone 3 (midfalling-rising), Tone 4 (high-falling), and Tone 5 (neutral). Furthermore, tonality of a Mandarin monosyllable is mainly characterized by the shape of its pitch contour. Pitch frequencies for each monosyllable are measured frame by frame and are in the form of $\{w_0(\frac{i}{N}), 0 \le i \le N\}$, where $N+1$ is the length of the pitch contour. The features for prosodic conversion are then extracted from the pitch contour by the orthogonal polynomial transform:

$$a_j = \frac{1}{N+1} \sum_{i=0}^{N} w_0(\frac{i}{N}) \cdot \Phi_j(\frac{i}{N}),\ j = 0, 1, 2, 3 \tag{5}$$

where $\Phi_j(\cdot)$ represent the first four discrete Legendre polynomials [7].

Our prosodic conversion system is based on vector quantization (VQ) and consists of two steps: a learning step and a conversion-synthesis step [3]. In the learning step, we trained the source speaker's codebook $\mathbf{C}^{(s)}$ and target speaker's codebook $\mathbf{C}^{(t)}$, and used them to generate a mapping codebook, $\mathbf{C} = \{\mathbf{c}_0, \mathbf{c}_1, \ldots, \mathbf{c}_{M-1}\}$, that specifies the correspondence between $\mathbf{C}^{(s)}$ and $\mathbf{C}^{(t)}$. In the conversion-synthesis step, the pitch contour of an isolated syllable was first obtained and its orthogonal transform coefficients form a vector, $\mathbf{a} = (a_0, a_1, a_2, a_3)^T$, which will be vector quantized by using the source speaker's codebook. The VQ encoder searches through the codebook $\mathbf{C}^{(s)}$ for the codevector that best matches $\mathbf{a}$ and then outputs its index k in binary format. Prosodic conversion was carried out by decoding the coefficient vector $\hat{\mathbf{a}} = \mathbf{c}_k$ using the mapping codebook $\mathbf{C}$. The converted pitch contour can be approximated as

$$\hat{w}_0(\frac{i}{N}) = \sum_{j=0}^{3} \hat{a}_j \cdot \phi_j(\frac{i}{N}),\ 0 \le i \le N \tag{6}$$

5. Experimental Results

Experiments were carried out to investigate the potential advantages of using spectral and prosodic conversion 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 training the mapping function and one for testing in our voice conversion experiment. Each set consisted of ten readings of 68 syllables, which are chosen to have combinations of various initial consonants and final vowels, and different tones. 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 speech was largely intelligible in sentences but often caused misunderstanding in syllables due to misarticulation of consonant phonemes and improper control of pitch contour.

Perceptual evaluations were made first to determine whether impaired speech samples sounded less intelligible than those converted using the GMM-based spectral conversion. The intelligibility of monosyllabic utterances was assessed by five native speakers of Mandarin Chinese. Table 1 shows the evaluation results in terms of average syllable identification accuracy. It was found that the spectral conversion resulted in more intelligible speech with an average identification score of 78.38%, compared to 60.88% for the impaired speech. The investigation further showed that an improvement of over 62% was obtained for syllables starting with affricate consonants. Cepstrum distances were also measured to test the validity of the proposed spectral conversion scheme. Table 2 shows the comparative results for ten Mandarin syllables, where the final vowel is /u/ and the initial consonant can be stop or affricate. To elaborate further, results of the 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 best results were seen in the affricates. 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 impaired speech, however, showed affricates that contained a complete stop. Our analyses revealed that the conversion softened the burst, removed their lower frequency energy and elevated the fricative portion to normal frequency ranges. The results are near perfect affricates. Experimental results also demonstrated the effectiveness of the VQ-based prosodic conversion for enhancing the quality of hearing-impaired speech. This is illustrated, for syllable /ti/ spoken with four different tones, in Figure 2 showing pitch contours of speech samples converted using a 4-bit VQ to encode the coefficient vector $\mathbf{a} = (a_0, a_1, a_2, a_3)^T$.

6. Conclusions

This study presents a novel means of exploiting voice conversion in the design of speaking aids for the hearing-impaired. Cepstral coefficients of the speech were converted from a hearing-impaired speaker to a normal speaker by a spectral conversion algorithm based on the Gaussian mixture model. By taking advantage of the tone structure of pitch contours in Mandarin speech, pitch information is orthogonally transformed and vector quantized to provide prosodic conversion. Evaluation by objective tests and listening tests shows that the proposed techniques greatly improve the quality and naturalness of the hearing-impaired Mandarin speech.

7. Acknowledgements

This study was supported by the National Science Council, Republic of China, under contract NSC 91-2614-E-009-001.

8. 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] Abe, M., Nakamura, S., Shikano, K., and Kuwabara, H., "Voice conversion through vector quantization", *Proc. ICASSP*, vol. 1:655-658, 1988.

[4] 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.

[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] 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.

[7] Chen, S. H. and Wang, Y. R., "Vector quantization of pitch information in Mandarin speech", *IEEE Trans. Communications*, vol. 38,:1317-1320, 1990.

Table 1: *Average syllable identification accuracy (%) for listening tests of impaired speech and converted speech.*

syllable	impaired speech	converted speech
affricate-vowel	30%	92.5%
fricative-vowel	55.83%	63.33%
stop-vowel	86.36%	87.27%
nasal-vowel	78.89%	85.56%
liquid-vowel	76%	80%
total average	60.88%	78.38%

Table 2: *Cepstrum distances before and after the spectral conversion.*

syllable	before conversion	after conversion
bu	12.72	3.82
pu	10.52	4.12
du	9.86	4.46
tu	10.34	4.40
gu	11.74	4.50
ku	11.25	4.46
zhu	13.93	6.58
chu	14.49	5.65
zu	15.24	6.38
cu	13.32	5.96

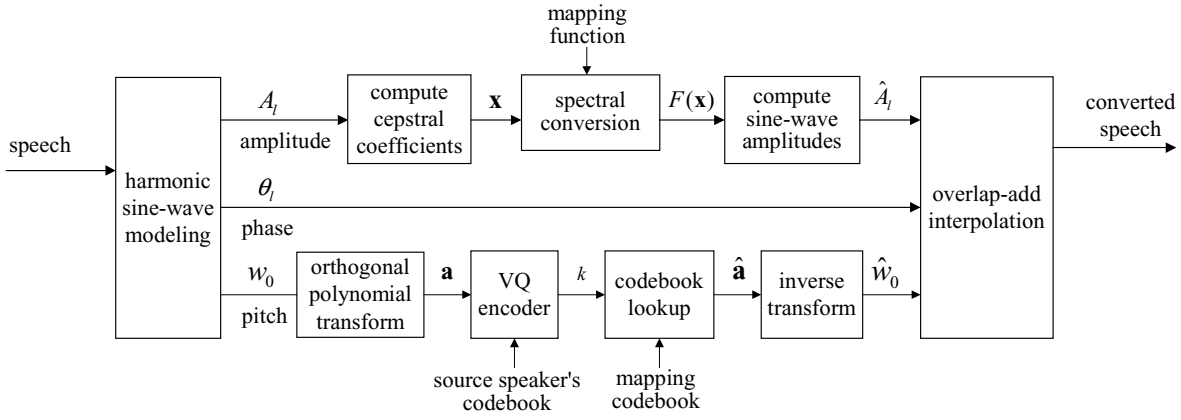

Figure 1: *Block diagram of the voice conversion system.*

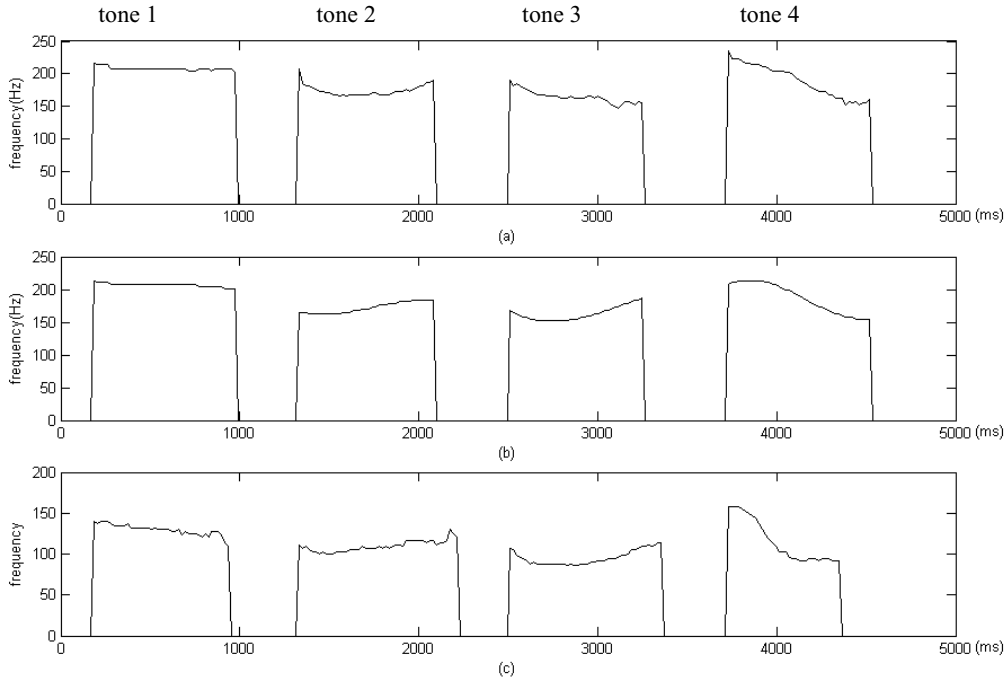

Figure 2: *Pitch contours for syllable /ti/ spoken with four different tones. (a) Impaired speech. (b) Converted speech. (c) Normal speech.*