Application of Speech Conversion to Alaryngeal Speech Enhancement

Ning Bi and Yingyong Qi

***Abstract*— Two existing speech conversion algorithms were modified and used to enhance alaryngeal speech. The modifications were aimed at reducing spectral distortion (bandwidth increase) in a vector-quantization (VQ) based system and the spectral discontinuity in a linear multivariate regression (LMR) based system. Spectral distortion was compensated for by formant enhancement using chirp z-transform and cepstral weighting. Spectral discontinuity was alleviated using overlapping clusters during the construction of conversion mapping function. The modified VQ and LMR algorithms were used to enhance alaryngeal speech. Results of perceptual evaluation indicated that listeners generally preferred to listen to the alaryngeal speech samples enhanced by the modified conversions over original samples.**

***Index Terms*—Speech enhancement, speech conversion, speech analysis and synthesis, vector quantization, linear multivariate regression**

I. Introduction

LARYNGEAL cancer may necessitate a total removal of the larynx, resulting in a fundamental change of speech production. For many alaryngeal individuals, voicing is mainly produced by setting surgically reconstructed tissues in the upper airway in vibration. Alaryageal speech sounds rough, hoarse, and creaky. A system that converts alaryngeal speech into normal speech could be useful to enhance communication for alaryngeal talkers [1], [2].

To enhance the quality of alaryngeal speech, Qi attempted replacing the voicing source of alaryngeal speech using a linear predictive coding (LPC) technique [1], [2]. There are two basic assumptions under these early studies: i) articulatory-based acoustic features of alaryngeal speech are not significantly modified by laryngectomy, and ii) vocal tract transfer functions of alaryngeal speech could be accurately determined using LPC analysis. These assumptions should be applicable to most alaryngeal speech because only the larynx is surgically removed during laryngectomy. In some special cases, however, these assumptions may not be valid. For example, the formant frequencies of alaryngeal speech may be significantly shifted upward due to the possible surgical shortening of the vocal tract. Larynx removal may also alter other articulatory behaviors because of the disrupted muscular support for the tongue. In these cases, both source- and articulation-related properties of alaryngeal speech need to be modified to achieve enhancement.

It has been documented that spectral conversion is a feasible technique for modifying articulation-related parameters of speech [3]–[9]. Spectral conversion was originally used for talker adaptation in speech recognition systems. The technique of spectral conversion was also used in normal voice conversion systems [4], [6], [7]. To accomplish voice conversion, the spectral space of an input talker was reduced to, and represented by an input codebook obtained using vector quantization (VQ) algorithms [10]. A mapping codebook that specifies the output vector of an input codeword was generated through a supervised learning procedure. Spectral conversion was accomplished by applying the mapping codebook to each input spectrum.

VQ-based spectral conversion method has two major sources of error/distortion. First, the reduction of a continuous spectral space into a discrete codebook introduces quantization noise, which inevitably creates a difference between a given spectrum and its corresponding codeword (representative spectrum) in the codebook. Second, under the cepstral representation, the codewords created by the VQ process typically are the means of a set of spectral clusters and, thus, have individual formant bandwidth larger than the original. In an effort to reduce quantization noise, Shikano *et al.* (1991) proposed a fuzzy vector quantization method in which an input spectrum was coded as a weighted interpolation of a set of codewords. This weighted interpolation has the potential to reduce quantization noise because the spectral space is now approximated by many interconnected lines between codewords rather than by a point grid of codewords. The weighted interpolation, however, increase further the bandwidth of the final coded spectrum.

A linear multivariate regression (LMR) approach for spectral conversion was used as an alternative to the VQ-based method [9]. In this approach, the spectral space of the input talker was partitioned by a few large clusters, and the spectra within each cluster was mapped linearly. The mapping matrix was obtained using procedures of least-square approximation. Because the mapping in a given region of the spectral space was continuous, the conversion distortions due to quantization and spectral averaging were minimized in a least square sense. The transitions between clusters in a connected speech, however, could be discontinuous resulting in audible clicks in the converted speech [9].

Manuscript received February 13, 1996; revised September 20, 1996. This work was supported in part by a grant from the National Institute of Deafness and Other Communication Disorders, DC01440, Analysis and Improvement of Alaryngeal Speech. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Douglas D. O'Shaughnessy.

N. Bi is with Qualcomm Inc., San Diego, CA 92121 USA.

Y. Qi is with the University of Arizona, Tucson, AZ 85721 USA (e-mail: yqi@u.arizona.edu).

Publisher Item Identifier S 1063-6676(97)01894-4.

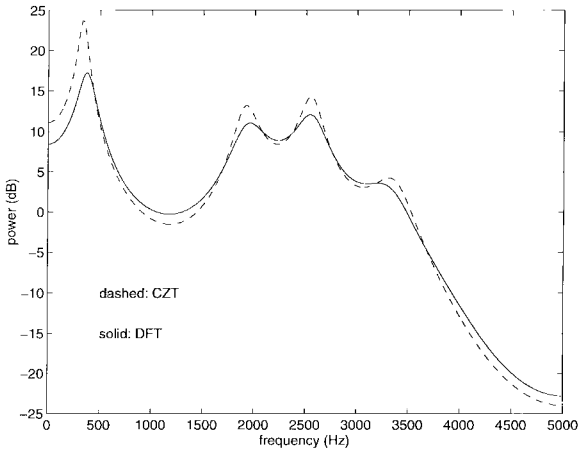


Fig. 1. Example of formant enhancement using the chirp z-transform.

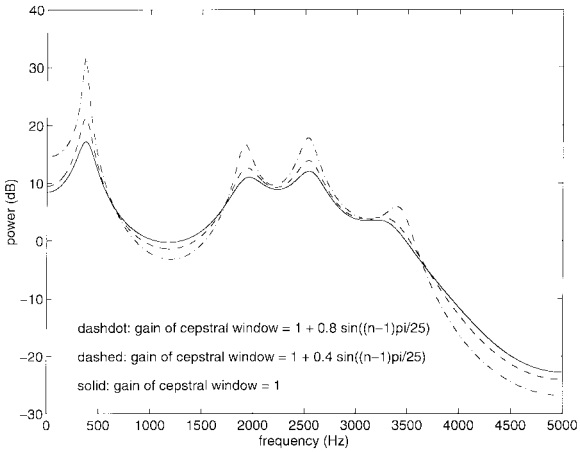


Fig. 2. Example of formant enhancement using cepstral weighting.

Despite of the problems of spectral averaging in VQ-based system and transition discontinuity in LMR-based system, it has been reported that the conversions were successful in that the converted speech is perceptually more close to the target than to the original speech [3]–[9]. Speech quality was not a major concern in these reported studies. However, the quality of speech would be the primary concern when using spectral conversion for speech enhancement.

The goal of this work is to improve the existing speech conversion methods and apply these speech conversion methods for the enhancement of alaryngeal speech. The specific objectives are:

- to modify the VQ-based method to reduce conversion distortions due to bandwidth increase;
- to modify the LMR-based method to reduce auditorily annoying, transitional discontinuities during speech conversion;
- to evaluate and compare the performance of VQ- and LMR-based systems;
- to determine if these modified spectral conversion methods can be used for alaryngeal speech enhancement.

II. Modifications of Spectral Conversion Methods

In this section, the modifications of VQ- and LMR-based spectral conversion methods are presented. These modifications are aimed at reducing the spectral distortion (bandwidth increase) in the VQ-based method and the spectral discontinuity in the LMR-based method.

A. Modification of VQ-Based Conversion Method

The bandwidth increase in the VQ-based speech conversion system is intrinsic to the algorithm of vector quantization. Vector quantization is an algorithm for choosing a limited set of codewords (spectra) that represent the whole spectral space of a given talker. Each codeword is essentially an averaging of a small cluster of spectra. The number of codewords and clusters is dependent on the algorithm and parameters chosen [10], [11]. Unfortunately, each codeword, being an averaged spectrum, tends to have a larger bandwidth than its constituents.

The bandwidth increase is also intrinsic to the VQ-based conversion mapping scheme, where the target spectrum is designated as the average of all the spectra projected from a given cluster in the input spectral space. A small cluster in the input spectral space might project divergently to a large area in the target spectral space. When the divergent projection occurs, the bandwidth of the target spectrum will be large.

Perceptually, speech synthesized with large bandwidth sounds ambiguous and unclear. Because spectral averaging cannot be avoided in the VQ-based spectral conversion system, our modified system included formant enhancement (bandwidth reduction of resonance/formant peaks) as part of the speech conversion process to compensate for the bandwidth increase. Formant enhancement was made after spectral conversion and before speech synthesis.

1) Formant Enhancement Using Chirp Z-Transform: One method to sharpen the spectral peaks/formants is to use the chirp z-transform [12]. The chirp z-transform allows for the evaluation of a transfer function on a contour that is not the unit circle. If the contour for computing spectral transfer function is located outside all poles of the transfer function and inside the unit circle, the bandwidth of the resulting transfer function will be reduced.

The z-transform of any sequence x_n is defined as

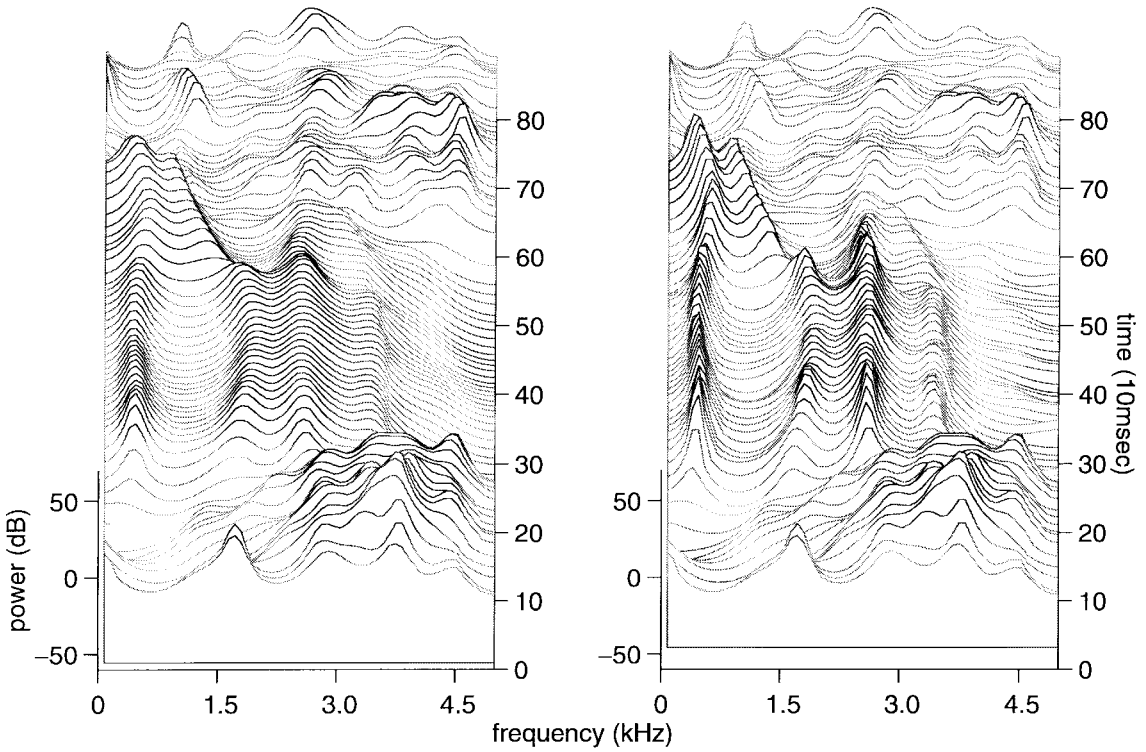
$$X(z) = \sum_{n=-\infty}^{\infty} x_n z^{-n}. \tag{1}$$

When $z = re^{jw}$, where r is an arbitrary complex number, (1) defines the chirp z-transform

$$X(jw) = \sum_{n=-\infty}^{\infty} x_n r^{-n} e^{-jnw}. \tag{2}$$

A special case of the chirp z-transform is when r is a constant and $|r| < 1$. It yields the z-transform of x_n on a circle with a radius $|r|$.

There are several ways to implement the chirp z-transform. One method is to multiply the LPC coefficients, a_i, by a factor, $a_i' = r^{-i} a_i$, and evaluate the adjusted polynomial on the unit circle [13]. The resulting spectrum will have sharper spectral peaks/formants than the original spectrum because the poles



(a) (b)

Fig. 3. Illustration of the use of formant enhancement during speech conversion. The conversion of the word sail was made: (a) by the conventional VQ-based method and (b) by the modified VQ-based method.

are effectively pushed out toward the unit circle. In order not to introduce extraneous variations during conversion, the magnitude of r should be a constant. It is difficult, however, to choose the magnitude of r *a priori*. If r is too large (close to the unit circle), it will not have significant formant sharpening. If r is too small (smaller than the magnitude of the largest pole of an LPC filter), it will make the LPC filter unstable.

An alternative is to implement the chirp z-transformation in the time domain. By substituting the system impulse response, h_n, with a weighted sequence, $r^{-n}h_n$, the transfer function of this system is evaluated on a circle inside the unit circle. To ensure the final synthesis filter is stable, the filter can be reestimated from linear predictive analysis of the weighted sequence using the autocorrelation method.

In our VQ-based conversion system, the chirp z-transform was implemented using the weighted impulse response. The magnitude of $r = 0.98$ was chosen based on the mapping codebook. It was the radius that set the upper-bound to the magnitude of all poles in the mapping codebook. The impulse response of new transfer function was obtained from the converted cepstrum [14]. An example of the converted spectrum before and after formant enhancement is shown in Fig. 1.

2) Formant Enhancement Using Cepstral Weighting: The formant enhancement effect using the chirp-z transform is limited by the magnitude of r. To enhance the formants further, the method of cepstral weighting was also applied [15].

The cepstrum for the vocal transfer function is a truncated segment of the whole cepstrum, obtained from the Taylor expansion of the log of LPC filter [16]. This windowing (truncation) operation is equivalent to a convolution in the frequency domain between the logarithmic spectrum of the original signal and the spectrum of the rectangular window. The spectrum of the rectangular window is characterized by a narrow mainlobe, but large sidelobes [17]. These sidelobes tend to smooth the resulting spectrum.

To enhance formants further, the rectangular window was replaced by a more rounded sine window, as follows:

$$w(n) = \begin{cases} 1 + h \sin\left[(n-1) * \dfrac{\pi}{L-1}\right], & \text{for } n = 1, 2, \cdots, L \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where h is a gain factor and was set to 0.4, the L is the window length and was set to 26. Because the sine window has smaller sidelobes than the rectangular window, it can reduce spectral smoothing to a certain extent. An example of formant enhancement using the sine window is shown in Fig. 2. An example of formant enhancement by applying both chirp z-transform ($r = 0.98$) and cepstral weighting ($h = 0.4$) is illustrated in Fig. 3.

B. Modification of LMR-Based Method

In the LMR-based approach, the spectral space was partitioned by a few large clusters and the spectrum within each cluster was mapped linearly [9]. The discontinuity in transitions between clusters in the LMR-based approach is, in part, caused by the use of a nonoverlapped clusters to

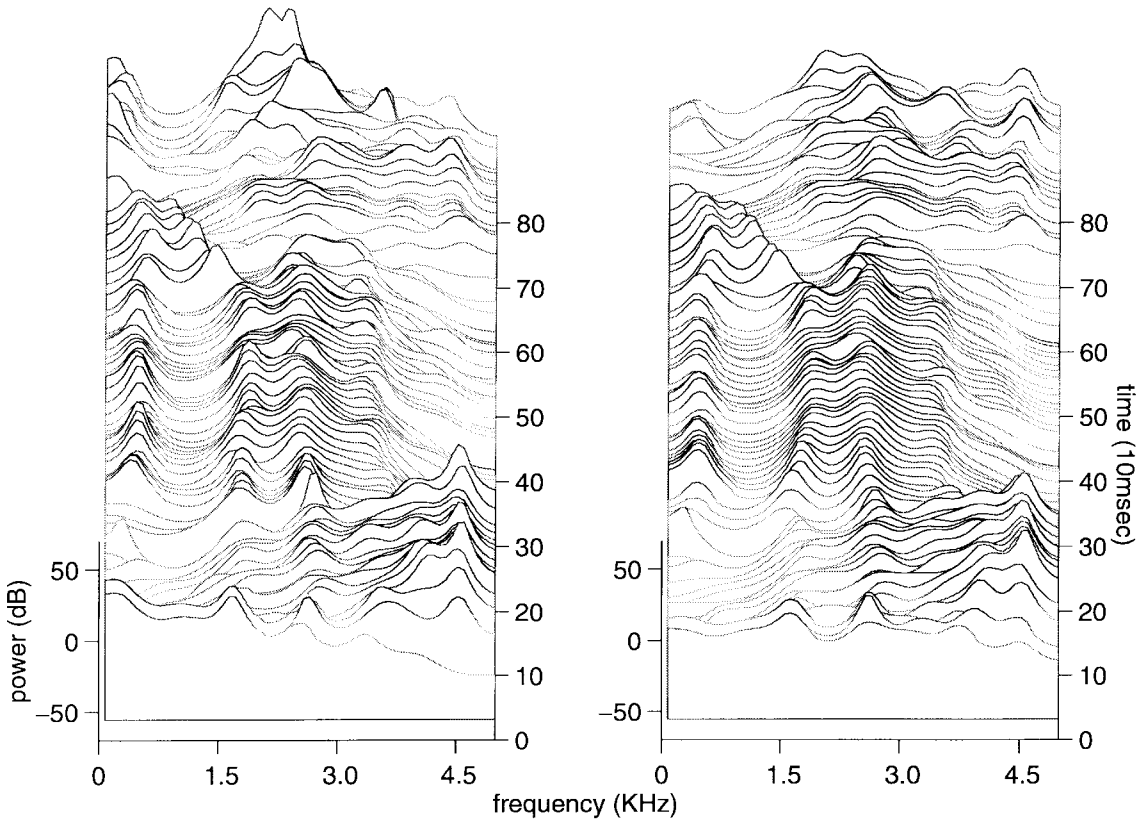


Fig. 4. Illustration of the use of overlapped subset training in speech conversion. The conversion of the word sail was made: (a) by conventional LMR-based method and (b) by the LMR-based method with overlapped subset training.

derive the LMR mapping matrix. Here, each mapping matrix is constrained only by samples of a given cluster and ignores the behavior of neighboring clusters. While each mapping matrix might serve its constituent cluster satisfactorily, neighboring mapping matrices may project toward different directions, resulting in spectral discontinuities during transitions between clusters.

In addition, some clusters may have a small number of elements. Thus, the mapping matrix may be constructed from an underdetermined rather than an overdetermined LMR problem when the number of elements in a given cluster is relatively small. The solution/pseudosolution (mapping matrix) of an underdetermined LMR problem can be problematic.

In our modified algorithm, an overlapped training method was used to reduce the spectral discontinuity [18], [19]. In this algorithm, overlapped clusters were used to obtain the LMR mapping matrix. The membership of a training sample, $\vec{v}$, is determined by the Euclidean distance, d_i, between the sample and the cluster centers, $\vec{c}_i$, $i \in 1, 2, \cdots, N$, where N is the total number of clusters. After reordering and renumbering the d_is according to their magnitudes, i.e., using $\tilde{d}_i \leq \tilde{d}_2 \leq \cdots \leq \tilde{d}_i \leq \cdots \leq \tilde{d}_N$ denote these distances, the training sample, $\vec{v}$, will participate in the training of cluster i if

$$D = \frac{\tilde{d}_1}{d_i} \tag{4}$$

is greater than a given threshold. The number of clusters that a training sample can join is limited to a maximum I. The overlapped area among neighboring subsets is controlled by the threshold. For example, when the threshold is 1, there will be no overlap. An example of using overlapped training in LMR-based spectral conversion is shown in Fig. 4, where the threshold is 0.75 and I is 6. It can be seen that the converted spectrogram is a more smooth function of time for the modified LMR-based conversion than for the original LMR-based conversion.

In summary, the advantages of using overlapped clusters during training are that:

- the mapping matrix of each cluster is constrained, to a certain extent, by samples of neighboring clusters so that continuity between transitions can be maintained;
- the size of training samples of each cluster is effectively increased so that the LMR mapping is likely to be an overdetermined problem as it should be.

III. System Implementation

The speech conversion system has four major components: speech analysis, voice source replacement, spectral conversion, and speech synthesis. The implementation of each component is described as follows.

A. Speech Analysis

Speech signals were analyzed to obtain LPC coefficients. Only the voiced segment of each utterance was analyzed. A signal segment (or frame) was considered to be voiced

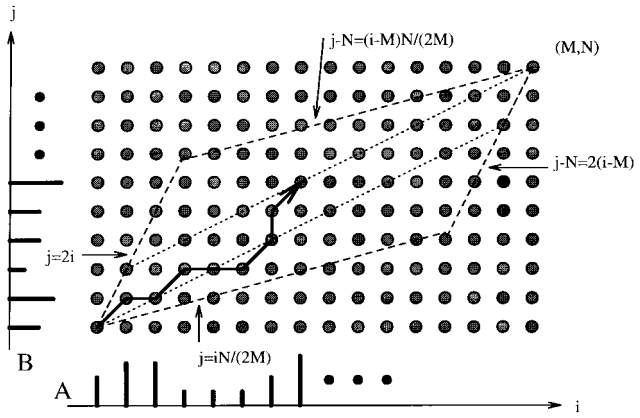


Fig. 5. Illustration of the parallelogram used in DTW matching.

when the fundamental period could be determined from the cepstral peaks of the signal [14]. The analysis window was 51.2 ms to include two or more periods for fundamental period determination. The fundamental period of an given speech segment was computed when its cepstral peak exceeded a preset threshold. The threshold of cepstral peak for alaryngeal speech was set to be half of that for normal speech due to the weak periodicity of alaryngeal speech [2]. The final periods were smoothed using a three-point median filter [20].

Fourteen LPC coefficients were computed for each voiced frame using the autocorrelation method [21]. Hamming window and pre-emphasis (0.98) were used in the LPC analysis. Frame length was set to 40 ms, and frame step-size was set to the current fundamental period. The LPC coefficients were transformed into 26 cepstral coefficients for spectral conversion and synthesis.

B. Voicing Source Replacement

The synthetic voicing excitation was generated based on the approximation of the LF-model [22]. The temporal parameters of the LF-model, t_e, t_p, and t_c, were defined as a constant proportion of the period. Amplitude, E_e was set based on the gain constant of the LPC filter.

C. Spectral Conversion

The spectral conversion rules between two talkers were built through a supervised learning procedure: an alaryngeal (input) talker, and a normal (target) talker, were asked to read the same list of words and sentences. The cepstra of these speech samples were computed every 5 ms. The computed spectral vectors of the same word or sentence were paired between the input and target talkers using the procedure of dynamic time warping (DTW) [23].

Because the duration of alaryngeal speech often is longer than that of normal speech, the warping region was adjusted adaptively to accommodate the spectral patterns to be matched. A warping parallelogram is illustrated in Fig. 5. Assuming M and N are the durations of two spectral patterns and $M > N$. the slope of the top and bottom sides of the warping parallelogram was set to $N/2M$ instead of a fixed 1/2 whereas the slope of the left and right sides was kept at 2 (the dotted lines). This adaptive modification of the warping region enabled the DTW algorithm to align most of the speech samples. The DTW total cost was used as a parameter to identify speech samples that time alignment was not possible. These samples were excluded from system training.

1) Implementation of VQ-Based Conversion System: The implementation of a VQ-based conversion system has two phases: the learning phase and the conversion-synthesis phase. In the learning phase, a mapping codebook that specifies the mapping function from the input spectral space to the target spectral space was generated. In the conversion-synthesis phase, speech signals were analyzed and, then, synthesized using the converted spectral transfer function.

During learning, the mapping codebook was generated from pairs of input and target spectral vectors. These spectral vector pairs were obtained using the analysis procedures described under Section III. Given the input and target vector pairs, the mapping codebook was obtained in the following three steps:

1) the codebook of input vectors (input codebook) was obtained using vector quantization;
2) the projections (target vectors) from a given input cluster were identified based on the pairing relations;
3) the average of these projections was designated to be the target codeword for the input cluster.

The sizes of input and target codebooks were set to 512. This process is illustrated in Fig. 6(a).

During conversion-synthesis, an input frame of signal was analyzed and its cepstral coefficients was obtained. The input codeword for the cepstral coefficients was identified and conversion was made based on the mapping codebook. To enhance the formants, the converted cepstral coefficients were weighted by the sine window before being transformed into system impulse response. The impulse response was weighted again by the sequence, $r^{-n}(r = 0.98)$ to enhance the formants further. A new set of LPC coefficients was re-estimated from this impulse response. A period of speech signal was then synthesized using these coefficients and the replaced excitation input. A block diagram of the conversion-synthesis process is illustrated in Fig. 6(b).

2) Implementation of LMR-Based Spectral Conversion: The implementation of LMR-based conversion also involves a learning phase and a conversion-synthesis phase. In the learning phase, a set of mapping matrices that specifies the mapping function from the input spectral space to the target spectral space was generated. In the conversion-synthesis phase, speech signals were analyzed and then synthesized using the converted spectral transfer function.

During learning, the mapping matrices were again generated from pairs of input and target spectral vectors. These vectors were obtained using the same supervised learning procedures as described in the previous section. Given the input and target vectors and the pairing relations, the mapping matrices were obtained as follows.

1) An input codebook of a few clusters (64) was obtained using vector quantization.
2) The projections of each input cluster were identified based on the pairing relations.

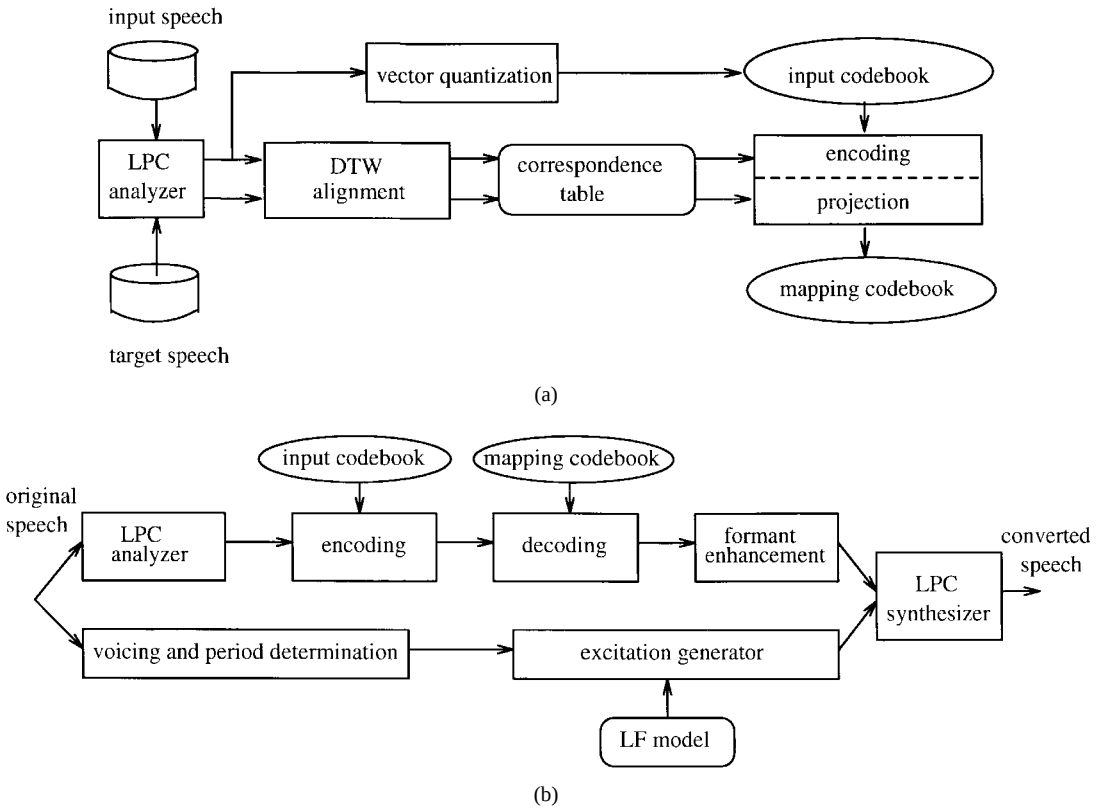


Fig. 6. (a) Block diagram of the learning process in the VQ-based conversion. (b) Block diagram of the conversion-synthesis process in the VQ-based conversion.

3) The vectors located on the edges of each subset also participated in the training of neighboring subsets. The threshold of normalized distance D was set to 0.75 and the parameter I was set to 6 [see (4)].
4) A mapping matrix, M, was computed using least-square approximations.

Let X denotes the input vectors in a given cluster and Y denote their projections in the target vector space. The least-square approximation proceeds with

$$M = YX^{\dagger} \tag{5}$$

where † denotes the pseudoinverse of X [24], [25] which is obtained as

$$X^{\dagger} = X^{T}(XX^{T})^{-1} \tag{6}$$

where ^{T} denotes the matrix transpose, and $^{-1}$ denotes the matrix inverse. This learning process is illustrated in Fig. 7(a).

In the conversion-synthesis phase of LMR-based system, an input spectrum is classified by the input codebook, and then is converted using the corresponding mapping matrix. A block diagram of the LMR-based system is shown in Fig. 7(b).

IV. Perceptual Evaluations

A. Subjects and Recordings

Normal speech samples were gathered from one male and one female talker. Alaryngeal speech samples were gathered from one male and one female tracheoesophageal takers. Both tracheoesophageal talkers were proficient and have used their method of alaryngeal speech for a minimum of one year. Both were referred to this project by the clinical speech pathologist responsible for their clinical speech rehabilitation treatment, and were rated average to above average in overall speech proficiency by their referring specialist.

Recordings were made of subjects producing 69 words and 25 sentences (C.I.D. Auditory Test W-1, California Consonant Test Items, and Competing-Sentence Test) at a comfortable level of pitch and loudness. The recordings (SONY, TCD-D3) were made in a quiet room with the recording microphone (ASTATIC, TM-80) placed about 5 cm from the mouth of each talker. The recorded words were digitized into a computer at a sampling frequency of 10 kHz (AT&T, DSP32-VME). The signal was passed through a low-pass filter (TTE, J73E) with a cut-off frequency of 4.5 kHz prior to digitization. All subjects read the C.I.D. Auditory Test W-1 and California Consonant Test Items twice, and the Competing-Sentence Test once. The first list of the recorded words and sentences were used for system learning, and the second list of the recorded words were used for conversion and perceptual evaluation.

B. Procedures of Perceptual Evaluation

Perceptual evaluations were made first to determine whether speech samples converted using the modified systems sounded more pleasant to the listeners than those converted using the unmodified systems. Fifty words produced by the normal male and female talkers were used for the evaluation. Conversions were made between the normal male and female talkers. A paired comparison procedure was used. Each word converted

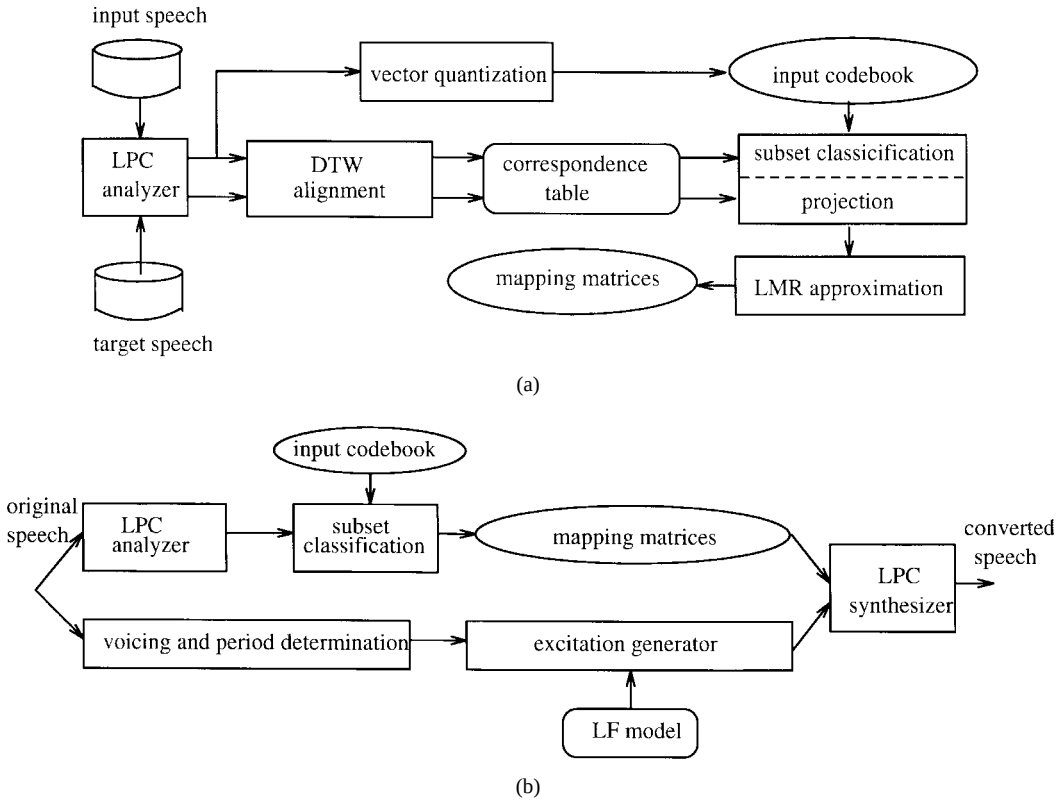


Fig. 7. (a) Block diagram of the learning process in the LMR-based conversion. (b) Block diagram of conversion-synthesis process in the LMR-based conversion.

using the modified system was paired up with the same word converted using the unmodified system. The order of the pair was random.

Twelve students at the University of Arizona provided the preference judgments. Each listener was allowed to listen to any pair of words as many times as needed before determining which word in the pair "sounded more natural or was more pleasant to listen to." Each listener also made preference judgments about the word pairs a second time on a different day. The order of the pairs in the list was rerandomized for the second presentation.

A paired comparison approach was also used to determine whether enhancement of alaryugeal speech was achieved using speech conversion systems. Six words (beach, drawbridge, inkwell, peep, sail, woodwork) produced by the alaryngeal talkers were selected for perceptual evaluation. These words were chosen because they provided a reasonably representative sampling of the vowel space.

Each word was synthesized under the following five conditions.

1) Only the voicing source was replaced.
2) Both the voicing source and the spectrum were replaced, and spectral conversion was made using the modified VQ-based conversion method.
3) Both the voicing source and the spectrum were replaced, and spectral conversion was made using the modified LMR-based conversion method.
4) Both the voicing source and the spectrum were replaced, and spectral conversion was made using the conventional VQ-based conversion method.
5) Both the voicing source and the spectrum were replaced, and spectral conversion was made using the conventional LMR-based conversion method.

Each original word and its 1–3 synthetic counterparts were paired in all possible combinations. Conditions 2 and 4, and 3 and 5, respectively, were also paired. All pairs were presented to the listeners. The order of the pairs in the presentation list was randomized. Perceptual judgments were made using the same procedure as described above.

C. Evaluation Results

The reliability of listeners was evaluated by calculating the percentages of agreement in preference judgments made by each listener in response to the repeated presentation of all word pairs (test-retest agreement). The responses of listeners exhibiting 50% or greater test-retest agreement in preference judgments were used to evaluate enhancement. Ten listeners achieved this arbitrarily established criteria.

Overall, 76% of the 2000 responses (2 talkers $\times$ 50 words $\times$ 10 listeners $\times$ 2 sessions) prefer words converted using the modified VQ system over the unmodified VQ system while 68% of the 2000 responses prefer words converted using the modified LMR system over the unmodified LMR system. Thus, moderate enhancement of speech produced by normal talkers was obtained using the modified conversion systems.

The listeners' judgments of preference made in response to words synthesized by different enhancement systems, and original word produced by the male, alaryngeal talker, are summarized in Table I. The data in Table I are the number and percentage of listeners preferring words synthesized under

TABLE I
NUMBER AND PERCENTAGE OF RESPONSES PREFERRING CONDITION OF WORD SPECIFIED IN THE FIRST COLUMN FOR THE MALE SUBJECT

word pair characteristics	alaryngeal speech	only source replaced	modified VQ-based	VQ-based	LMR-based
only source replaced	68% 82/120				
modified VQ-based	82% 98/120	87% 104/120		64% 77/120	
modified LMR-based	80% 96/120	85% 102/120	50% 60/120		62% 74/120

TABLE II
NUMBER AND PERCENTAGE OF RESPONSES PREFERRING CONDITION OF WORD SPECIFIED IN THE FIRST COLUMN FOR THE FEMALE SUBJECT

word pair characteristics	alaryngeal speech	only source replaced
only source replaced	96% 115/120	
modified VQ-based	90% 108/120	43% 52/120

conditions described in the first column. The total number of responses for each comparison is 120 (6 words $\times$ 10 subjects $\times$ 2 sessions).

Based on a binomial distribution table [26], these data reveal a significant ($p < 0.01$), clear overall preference by the listeners for the synthesized versions of words, demonstrating that enhancement of speech produced by this male laryngectomized talker, was accomplished using speech analysis-synthesis methods with or without spectral conversion.

The data in Table I also revealed the impact of spectral conversion. Listeners preferred converted words over the words synthesized by replacing voicing source only. As expected, both the modified VQ- and LMR-based speech conversion approaches achieved better performances than the conventional systems. The modified LMR-based method and the VQ-based method had comparable performance.

For the female alaryngeal talker, speech enhanced by the LPC analysis-synthesis method had the highest scores (see Table II). Listeners almost unanimously preferred synthesized version of words over the originals. Listeners also preferred the speech samples synthesized by LPC analysis-synthesis without spectral conversion.

These results indicated that speech conversion would be useful for alaryngeal talkers with articulatory deficits. The speech conversion would not be necessary when articulatory deficits are minimal. A voice source replacement alone would provide a significant enhancement [2].

V. DISCUSSIONS AND CONCLUSIONS

Formant enhancement described in the modified VQ-based algorithm could also be applied to the LMR-based system because the LMR least-square mapping also introduces some spectral averaging. The magnitude of averaging in the LMR-based system, however, is much smaller than that in the VQ-based system. Hence, formant enhancement was not implemented in the modified LMR-based system to focus attention on the overlapped training and its effect.

The cepstrum-based, fundamental period determination algorithm may not work well for signal segment that has weak periodicity. For example, it may misclassify some transitional voiced segment as unvoiced. This type of misclassification, however, is not expected to influence the results significantly because the quality of voiced segment of speech is determined primarily by those segments that carry an appreciable amount of energy [27].

The increase of perceptual evaluation scores due to system modifications is larger for the normal speech than for the alaryngeal speech. This difference may be attributed, in part, to the difference of data set used. Six of the 50 words used for normal speech comparison were used in alaryngeal speech comparison. In addition, the improvement of system modifications might be difficult to observe when the overall quality of the speech samples used are very poor. A more comprehensive evaluation may be needed using a large database of alaryngeal talkers. Unfortunately, we could only locate one male talker, that has articulatory deficit in his production of alaryngeal speech.

In conclusion, the original VQ- and LMR-based spectral conversion methods were modified. The modifications were aimed at reducing the spectral distortion in the VQ-based method and the spectral discontinuity in the LMR-based method. The modified systems were used for alaryngeal speech enhancement. Perceptual evaluations based on a limited data set were completed to determine if enhancement could be accomplished using these modified speech conversion methods. Results of perceptual evaluations indicated that listeners generally preferred the output of the modified algorithms. The enhancement achieved by the modified LMR-based approach was comparable to that of the modified VQ-based approach. Results of perceptual evaluations also revealed that speech conversion techniques were more effective on alaryngeal speech with articulatory deficits when comparing to enhancement achieved by voice source replacement alone.

REFERENCES

[1] Y. Qi, "Replacing tracheoesophageal voicing sources using LPC synthesis," *J. Acoust. Soc. Amer.*, vol. 88, pp. 1228–1235, 1990.
[2] Y. Qi, B. Weinberg, and N. Bi, "Enhancement of female esophageal and tracheoesophageal speech," *J. Acoust. Soc. Amer.*, vol. 97, pp. 2461–2465, 1995.
[3] K. Shikano, K. Lee, and R. Reddy, "Speaker, adaptation through vector quantization," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Tokyo, Japan, 1986, vol. ICASSP-86, pp. 2643–2646.
[4] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. IEEE Int. Conf. Acoutics, Speech, and Signal Processing*,New York, 1988, vol. ICASSP-88, pp. 655–658.
[5] S. Nakamura and K. Shikano, "Speaker adaptation applied to HMM and neural networks," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1989, vol. ICASSP-89, pp. 89–92.
[6] M. Abe, K. Shikano, and H. Kuwabara, "Cross-language voice conversion," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1990, vol. ICASSP-90, pp. 345–348.
[7] M. Abe, "A segment-based approach to voice conversion," in *Proc. of IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Toronto, Canada, 1991, vol. ICASSP-91, pp. 765–768.

[8] K. Shikano, S. Nakamura, and M. Abe, "Speaker adaptation and voice conversion by codebook mapping," in *Proc. IEEE Int. Symp. Circuits and Systems*, 1991, vol. 1, pp. 594–597.
[9] H. Valbret, E. Moulines, and J. Tubach, "Voice transformation using PSOLA technique," *Speech Commun.*, vol. 11, pp. 175–187, 1992.
[10] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, Jan. 1980.
[11] L. Rabiner, S. Levinson, and M. Sondhi, "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition," *Bell System Tech. J.*, vol. 62, pp. 1075–1105, 1983.
[12] L. Rabiner, R. Schafer, and C. Rader, "The chirp z-transform algorithm," *IEEE Trans. Audio Electroacoust.*, vol. AU-17, pp. 86–92, 1969.
[13] S. McCandless, "An algorithm for automatic formant extraction using linear predictive apectra," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 135–141, 1974.
[14] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals.* Englewood Cliffs, NJ: Prentice-Hall, 1978.
[15] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ: Prentice-Hall, 1993.
[16] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 381–391, Oct. 1976.
[17] A. Oppenheim and R. Schafer, *Discrete-Time Signal Processing.* Englewood Cliffs, NJ: Prentice-Hall, 1989.
[18] H. Matsumoto and H. Inoue, "A piecewise linear spectral mapping for supervised speaker, adaptation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1992, vol. ICASSP-92, pp. 449–452.
[19] N. Iwahashi and Y. Sagisaka, "Speech spectrum conversion based on speaker, interpolation and multi-functional representation with weighting by radial basis function networks," *Speech Commun.*, vol. 16, pp. 139–151, 1995.
[20] L. Rabiner and M. Sambur, "Application of an LPC distance measure to the voiced-unvoiced-silence detection problem," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 338–343, Aug. 1977.
[21] J. Markel and A. Gray, *Linear Prediction of Speech.* New York: Springer-Verlag, 1976.
[22] Y. Qi and N. Bi, "A simplified approximation of the four-parameter LF model of voice source," *J. Acoust. Soc. Amer.*, vol. 96, pp. 1182–1185, 1994.
[23] L. Rabiner, A. Rosenberg, and S. Levison, "Considerations in dynamic time warping algorithms for discrete word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 575–582, 1978.
[24] T. Kohonen, *Associative Memory*, 1st ed. New York: Springer-Verlag, 1977.
[25] ———, *Self-Organization and Associative Memory*, 3rd ed. New York: Springer-Verlag, 1989.
[26] W. MacKinnon, "Table for both the sign test and distribution free confidence intervals of the median for sample sizes to 1000," *J. Amer. Stat. Assoc.*, vol. 59, pp. 935–956, 1964.
[27] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, 2nd ed. New York: Springer-Verlag, 1972.

Ning Bi received the B.S. degree in physics from the Peking University, China, in 1983, the M.S. degree in bioacoustics from the Institute of Zoology, Chinese Academy of Sciences, Beijing, in 1986, and the Ph.D. degree in speech science from the University of Arizona, Tucson, in 1995.

He worked as an Assistant Research Fellow in the speech recognition laboratory of the Institute of Acoustics, Chinese Academy of Sciences, from 1986 to 1991. He was a co-inventor of a real-time speech recognition system that received the international prize of TEC-88, Grenoble, France. From 1991 to 1995, he was a Research Assistant with the Department of Speech and Hearing Sciences, University of Arizona, and worked on alaryngeal speech enhancement. In the summers of 1994 and 1995, he worked on speech recognition algorithms and communication network protocols as a research internship at Hewlett-Packard Labortories, Palo Alto, CA. He is currently a Senior Engineer on speech signal processing at the Qualcomm Inc., San Diego, CA. His research interests include speech coding, recognition, conversion, and enhancement.

Yingyong Qi received the B.S. degree in physics at the University of Science and Technology of China, Hefei, in 1983, and the M.S. degree in acoustics at the Institute of Acoustics, Chinese Academy of Sciences, Beijing, in 1985. He received the Ph.D. degree in speech science from the Ohio State University, Columbus, in 1989, and a second Ph.D. degree in electrical engineering from the University of Arizona, Tucson, in 1993.

From January 1996 to January 1997, he has been a Visiting Scientist at the Research Laboratory of Electronics, Massachusetts Institute of Technology. He has been an Assistant/Associate Professor with the Department of Speech and Hearing Sciences, University of Arizona, since 1989. His major research interests include speech acoustics, alaryngeal speech enhancement, digital speech and image processing, and pattern recognition.

Dr. Qi is a member of the Acoustical Society of America and the first recipient of the Dennis Klatt Memorial Award given by the Acoustical Society of America.