International Journal of Electronics Communication and Computer Engineering
Volume 5, Issue 2, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

High Quality Voice Morphing using Linear Predictive Coding Method (LPC)

Rohini S. Dhorje
College of Engineering & Research, Wagholi, Pune
Email: rohini_dhorje@rediffmail.com

Prof. P. R. Badadapure
College of Engineering & Research, Wagholi, Pune
Email: badadapurepr@rediffmail.com

Abstract **– Voice morphing is a technique for modifying a source speaker's speech to sound as if it was spoken by some designated target speaker. One of the most recent models for voice conversion is the classical LPC analysis-synthesis model combined with GMM, which aims to separate information from excitation and vocal tract and to learn the transformation rules with statistical methods. However, it does not work well as it is supposed to be due to the inaccuracy of the extracted feature information as well as the overly-smoothed spectral converted by traditional GMM. In this paper, we propose a novel method to solve the problem which is based on the technique of the separation of glottal waveforms and the prediction of the excitations. The final result shows that not only are the transformed vocal tract parameters matching the target one better, but also is the high quality of the synthesized speech preserved.**

Keywords **– Voice Morphing, GMM, LPC, Feature Extraction, Glottal Waveform Separation.**

I. INTRODUCTION

There are many applications of voice morphing including customizing voices for text to speech (TTS) systems, transforming voice in adverts to sound like that of a well-known celebrity, and improving the intelligibility of abnormal speech uttered by a person with a speech problem.

In general, almost all the voice morphing systems consist of two stages: training and transforming, of which the core process is the transformation of the spectral envelope of thesource speaker to match that of the target speaker. In order to implement the personality transformation, two problems need to be considered: how to convert the vocal tract related feature parameters as well as excitation information. Until recently, many of previous published VC approaches have been centered on vocal tract mapping whose features areparameterized by some related LPC parameters, i.e., LSFs [1]-[5]. However it has already been reported that some kinds of transformation need to be applied to excitation signals in order to achieve high quality transformation [6][14][15]. Furthermore, the converted speech is often suffered from the degraded quality caused by traditional GMM-Based mapping method due to its overly-smoothed converted spectral problem [5].

In order to achieve high quality converted speech, the main problems mention above need to be solved. In this paper, we present a novel voice morphing system which separates excitation signal from the speech waveform in order to transform vocal tract parameters precisely. Also, a new strategy to convert excitation information based on residual prediction technique is proposed. The remainder of this paper is organized as follows.

First, an overview of our voice conversion framework is given in section 2, followed by the detailed description of proposed technique in section 3. In section 4, the performance of the proposed system with these new techniques integrated is evaluated. Finally, overall conclusions are presented in section 5.

II. OVERVIEW OF THE PROPOSED SYSTEM

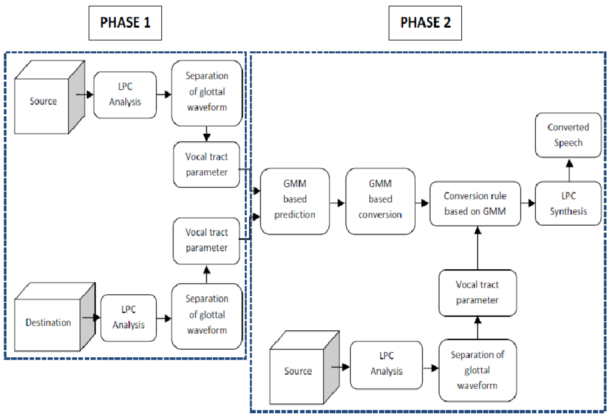


Fig.1. Block diagram of the proposed system

The framework of our voice conversion system is shown in Figure 1. The system consists of two procedures: the training procedure and the transforming procedure. In the training stage (phase1), voices from source and target speaker were firstly segmented in frames of two pitch period lengths and then an analysis based on LPC model was performed to extract vocal tract feature vectors to be transformed. In this work, a glottal waveform separation technique was proposed leading to achieving much more precise vocaltract parameters than the baseline system of which LSF parameters were estimated. A good alignment between source and target features was required to train the system, so dynamic time warping (DTW) was applied in preprocessing step. The basic spectral conversion rule is essentially equivalent to that proposed by Stylianou et al. [4], however, in order to achieve an effective personality change it is also needed to modify the glottal excitation characteristics of the source speaker to match the target one, so a prediction rule was trained on the aspects of the excitation signals of the target speaker as described in [17]. In the transforming stage (phase2), the extracted source vocal tract features were modified based on the conversion rule from the training stage, meanwhile, converted excitation signals were obtained by predicting from the transformed vocal tract features based on the prediction rule. Finally, continuous waveforms were obtained by synthesizing all these parameters in LPC synthesis model.

International Journal of Electronics Communication and Computer Engineering
Volume 5, Issue 2, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

III. Detailed Algorithm

A. Glottal waveform separation algorithm

According to the LP algorithm [19], an effective speech production model (for voiced speech) is given in Figure 2

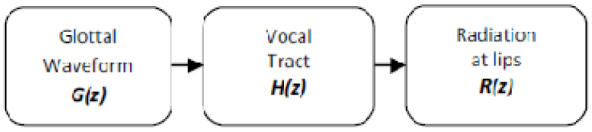


Fig.2. Speech production model

Where $S(z)$ represents acoustic speech waveform, $G(z)$ represents glottal waveform shaping, $V(z)$ models the vocal tract configuration and $R(z)$ represents the radiation at the lips which can be modeled as an effect of differential operator. So $G'(z)$, which is named for glottal derivative, can be derived as the product of $G(z)$ and $R(z)$

$$S(z) = G'(z)V(z) = G(z)V(z)R(z) \quad (1)$$

Given this model assumption, it was obviously that we could directly obtain the explicit representation of vocal tract by inverse filtering $S(z)$ with $G'(z)$, i.e.,

$$V(z)=S(z) / G'(z) \quad (2)$$

Unfortunately, it isn't the fact. It means that the solution to $V(z)$ mention above doesn't work well due to the existence of the disturbance from $G'(z)$. And it is believed that more precise vocal tract parameters could be obtained if the effect of glottal derivative was removed.

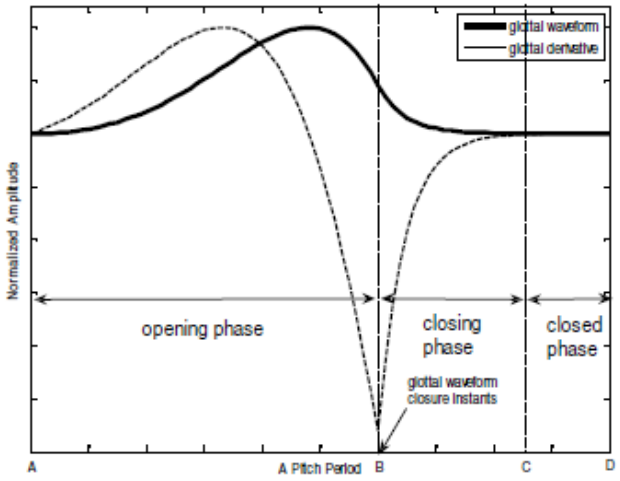


Fig.3. Glottal waveform and glottal derivative

Figure 3 shows an ideal glottal waveform as well as glottal derivative in a pitch period. Note that the amplitude of the glottal waveform starting from the glottal closure instants (GCI) to the end of the pitch period is decreased monotonically, i.e., during the closing phase and the closed phase of the glottal waveform, the interaction with vocal tract is decreased monotonically, where linear prediction (LP) analysis can be performed to model the vocal tract almost exclusively since glottal contribution is minimal. This conclusion [19] can be interpreted by rewritten the formula (2) as

$$S(z) = G'(z)V(z) = \frac{\sum_{j=1}^{q} b_j z^{-j}}{\sum_{i=0}^{p} a_i z^{-i}} \quad (3)$$

Take this z-transform into time domain, we have

$$S(n) = -\sum_{i=1}^{p} a_i S(n-i) + g(n) \quad (4)$$

Where $g(n)$ denotes the glottal excitation in time domain, and it has the following expression

$$g(n) = b_j \sum_{j=1}^{q} \delta(n-j) \quad (5)$$

Note that if the time index n goes into the region where $g(n) \approx 0$, i.e., the closing phase and the closedphase, then equation (4) will reduce to

$$S(n) \approx -\sum_{i=1}^{p} a_i S(n-i) \quad (6)$$

This is the generalized LPC formula which we used in common speech analysis process which implicates that it is advantageous to obtain an accurate estimation of the vocal tract in the closing phase or closed phase of the glottal flow. However, the traditional LPC analysis is actually performed during the whole pitch period, not the closing phase. Now the problem comes to how to locate glottal closure instants (GCI). The algorithm presented in this paper borrows the main principle from [11] which proposed a simplified method to obtain the best possible estimation of GCI. When GCI is determined, we can obtain a precise vocal tract response in the closing or closed phase of glottal flow by deconvolution of the speech signal using traditional LP technique.

B. Excitation signal prediction

As mention in section 1, in order to achieve high quality converted speech, excitation information needs to be taken into account. Previous work on excitation signal conversion mainly focuses on the modification on the source speaker excitation signal to match with the target one [16]. This idea is similar to the vocal tract conversion. However, there always exist overly smoothed problems with the GMM-Based conversion rule which will lead to degraded synthesized speech quality.

In this section, excitation signals are predicted from the spectral parameters in contrast to directly transforming them. In order to predict the excitations from their LSF parameters, the assumption that the excitations are completely uncorrelated with the spectral envelopes is broken. For a particular speaker it is expected that the excitations corresponding to phones of acoustically similar classes are similar and predictable. According to figure 1, in the training stage, an analysis was performed to extract vocal tract parameters with their corresponding glottal excitations, both of which were stored in matrices respectively where the number of the rows was equal to the number of the analyzed frames. Then a GMM model was used to fit the probability of the target speaker's vocal tract parameters as

$$P_{GMM}(\mathbf{y}) = \sum_{q=1}^{Q} \omega_q N(\mathbf{y}; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) \quad (7)$$

where **y** denote the vocal tract vectors, the number of the mixture component is Q, and the weight of the *q*th mixture is $q\omega$, which satisfies

$$\sum_{q-1}^{Q} \omega_q = 1.$$

Given the frame number *t* and the mixture component index *q*, the probability of *t* **y** belonging to the *q* th mixture is given by

$$h_q(\mathbf{y}_t) = \frac{\omega_q N(\mathbf{y}_t; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)}{\sum_{q=1}^{Q} \omega_q N(\mathbf{y}_t; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)} \qquad (8)$$

Once the posterior probability is calculated, the prediction rule of the excitation signals can be populated. We use the stored matrices consisting of vocal tract vectors as well as their corresponding excitations to build the rule. Let $\mathbf{M_t}$ denotes the target magnitude of the excitation signal of frame *t*, then the magnitude of the *q* th component of the model is

$$\mathbf{m}_q = \sum_{t-1}^{T} \mathbf{M}_t \frac{h_q(\mathbf{y}_t)}{\sum_{t=1}^{T} h_q(\mathbf{y}_t)} \qquad (9)$$

where *q* =1,2,….,Q.

In the transforming stage, the posterior likelihoods of the incoming converted target vocal tract parameters were first computed through GMM mapping function, and then they were used as weights in predicting the excitations' magnitudes by a weighted mean scheme as following

$$\hat{\mathbf{M}}_t = \sum_{q=1}^{Q} \mathbf{m}_q h_q(\hat{\mathbf{y}}_t) \qquad (10)$$

Where $t\Lambda\mathbf{M}$ denotes the predicted excitation magnitude of frame *t* .

IV. Evaluation

Both objective and subjective experiments were performed to evaluate the performance of the proposed method. The speech corpus for this study consists of 250 parallel utterances of Mandarin Chinese spoken by a male and a female which we refer to M and F respectively. These data are sampled at 16 kHz and quantized for 16 bit per sample in a quiet environment. 180 utterances had been used for training and the remaining for the test. Note that in our test experiments, both glottal flow separation and excitation prediction technique are used in voiced frames. It means that for unvoiced frames, the source information is simply copied to the converted speech.

A. Spectral comparison

The comparison of the spectral envelope of an arbitrary frame obtained by the traditional and the proposed system is given in Fig 4and Fig 5. Note that the spectral of the same frame obtained by the two systems look different from each other, which shows that, according to the proposed system, not only has the formant structure of source speech been transformed to more closely match the target one, but also the spectral details been maintained successfully.

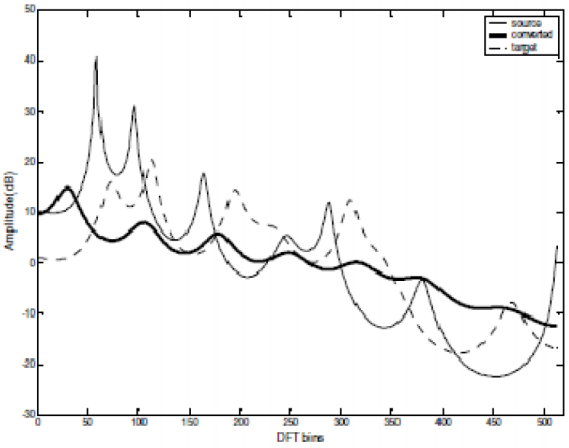


Fig.4. Spectral converted by traditional system

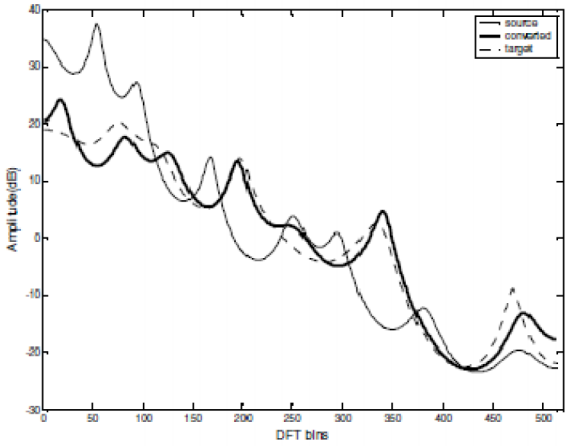


Fig.5. Spectral converted by proposed system

B. Time domain waveform comparison

The comparison of the time domain waveform is given in figure 6, which obviously shows that the proposed system performs better than the traditional one, preserving more details in the reconstructed waveform.

C. Signal to Noise Ratio (SNR) evaluation

A log distortion measure in time domain is used to evaluate the objective performance of the voice morphing system, which is defined as

$$SNR = 10\log 10\left(\frac{\sum s(t)^2}{\sum (s(t) - s_c(t))^2}\right) \qquad (11)$$

where *s*(*t*) and *s* (*t*) *c* denote target voice and converted voice separately. Table I shows the result of the comparison of the distortion of the traditional morphing system and the proposed system.

Table I. Signal to noise ratio of the systems

System	Traditional	Proposed
SNR(dB)	2.0471	2.9865

D. ABX evaluation

In order to test the overall subjective quality, listening tests were conducted to assess the perceptual accuracy of the proposed system. A so called ABX test was conducted whereas 5 listeners were asked to judge whether an utterance X sounded closer to utterance A or B in terms of

speaker identity where X was referred to the converted speech and A and B were source and target speech respectively. Table II gives the percentage of the converted utterances that were labeled as closer to the source one or target one or neither of them.

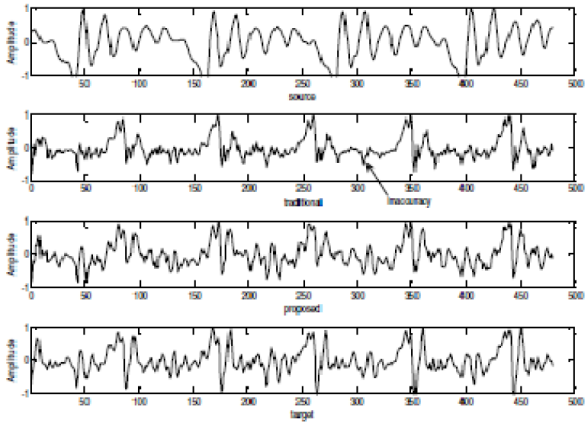


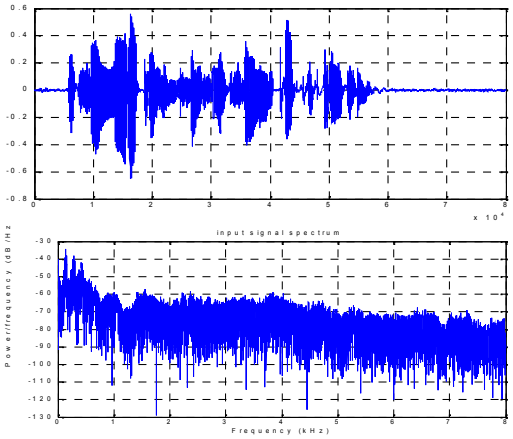
Fig.6. Comparison of the waveform. The horizontal shows the time axis, and the vertical shows the frequency axis with normalized amplitude (top: source, bottom: target, the one below top: traditional, the one above bottom: proposed)

Table II. Results of the ABX test

	Source	Target	Neither
Traditional	20.1%	33.6%	46.3%
Proposed	10.5%	63.4%	26.1%

E. Results

Both female and male speech signals were processed and the speech morphing algorithm was applied on these two signals where the female is the source signal and the male is the target one. The time domain of one frame of the source, target and morphed signals is shown in Figure 7. It is noticed here that although the morphing signal has approximately the same period as that of the target one but still there are noticeable differences in shape, duration and energy distribution. The pitch period of the whole signal source, target and morphed) was monitored as illustrated in Figure VII Obviously, pitch of the morphed signal is similar to that of the target one.



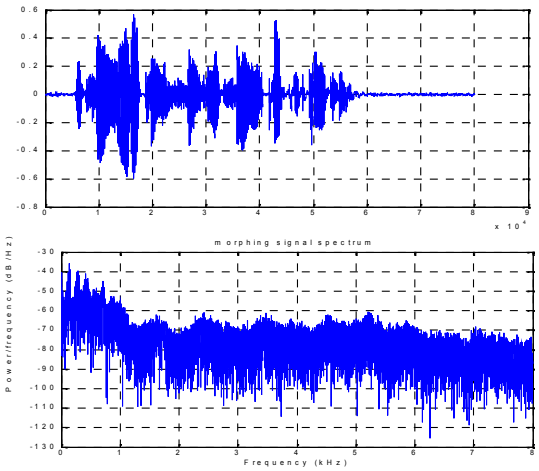


Fig.7. Source signal, source spectrum, target signal, target spectrum

V. Conclusion

This paper presents a novel method which is based on the technique of the separation of glottal waveforms and the prediction of the transformed residuals for precise voice conversion. The final result shows that not only are the transformed vocal tract parameters matching the target one better, but also are the target personalities preserved. Although the enhancements described in this paper give a substantial improvement, there is still distortion remained which makes the audio quality depressive and the future work will therefore focus on it.

Acknowledgment

Our sincere thanks to Prof. S.L.Lahudkar, Electronics & Telecommunication Department, ICOER, Pune, for contributing towards the development of the proposed system by providing us within formation regarding feature extraction and the newer solutions being deployed to enable high quality morphing.

The authors would also like to thank Prof. P.S. Topannavar, Department of Electronics & Telecommunication, ICOER, Pune, for his prompt guidance regarding the project.

References

[1] Abe M., Nakamura S., Shikano K. and Kuwabara H., “Voice conversion through vector quantization”, ICASSP, 1988:655~658

[2] Baudoin G., Stylianou Y., “On the transformation of the speech spectrum for voice conversion”, ICSLP’96, Philadephia, October 1996, 2:1405-1408

[3] Kain A. and Macon M., “Spectral voice conversion for text to speech synthesis”, ICASSP, 1998-05, 1:285-288

[4] Stylianou Y. and Cappe O., “A system for voice conversion based on probabilistic classification and a harmonic plus noisemodel”, ICASSP, 1998, Seattle, Washington, USA , pp.281-284 ,

[5] Hui Ye and Steve Young, “High quality voice morphing”, ICASSP, 2004, Montreal, Canada

[6] Levent M., Arslan and David Talkin, "Voice conversion bycodebook mapping of line spectral frequencies and excitationspectrum", Eurospeech, .pp. 1347-1350, 1997

[7] Tomoki Toda, Hiroshi Saruwatari and Kiyohiro Shikano, "STRAIGHT based voice conversion algorithm based on Gaussian mixture model", ICSLP, Beijing, China, pp. 279-282, 2000[8] Valbert H ., Moulines E . and J.P. tubach,"Voice transformation using PSOLA techniques", Speech Communication, Vol.11, pp.175-187,1992

[9] Takigi T. and Kuwabara H., "Acoustic parameters of voiceindividuality and voice quality control by analysis-synthesismethod", Speech Communication, Vol.10, pp. 491-495,1991

[10] Lee K., "A new voice transformation method based on bothlinear and nonlinear prediction analysis", ICSLP ,1996:1401~1404

[11] Elliot Moore, Mark Clements, "Algorithm for automaticglottal waveform estimation without the reliance on precise glottalinformation", ICASSP 2004

[12] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", J. R. Stat.Soc. B ,vol.39, pp. 1-22

[13] Kuwabare H. and Sagisaka Y., "Acoustic characteristics of speaker individuality: Control and conversion", vol.16,pp. 165-173

14] Nam I. H., "Voice personality transformation", Ph.D. Thesis,Electrical Engineering, Rensselaer Polytechnic Institue, Troy ,NewYork ,1991

[15] Weber F., Manganaro L., Peskin B. and E. Shriberg, "Using prosodic and lexical information for speaker identification", ICASSP, 2002

[16] Duxans H., Bonafonte A., "Residual conversion versusprediction on voice morphing system", ICASSP, 2006

[17] Ki Seung Lee, "Statistical approach for voice personality transformation", IEEE Trans on audio, speech and language processing, vol. 15, no. 2, 2007

[18] Kun Liu, etc., "high quality voice conversion throughcombining modified GMM and formant mapping for Mandarin", ICDT, 2007

[19] Xuedong Huang, etc., "Spoken Language Processing: AGuide to Theory, Algorithm and System Development", ISBN:0-13-022616-5