ISSN:2231-0711

High Quality Voice Morphing based on Residual Prediction Technique

Rohini S. Dhorje, Prof. P.R. Badadapure

Imperial College of Engineering and Research,
Wagholi, Pune, India.

***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. Also, a new strategy to convert excitation information based on residual prediction technique is proposed. 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*—Morphing, GMM, Extraction, LPC, analysis, synthesis , Glottal Waveform Separation.**

I. INTRODUCTION

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 the source 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 are parameterized 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] [7][8].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.

This project 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. In generalized LPC formula which we used in common speech analysis process which implicates that it is advantageous to obtain an accurate estimation of vocal tract in closing phase or closed phase of glottal flow. However the traditional LPC analysis is actually performed during the whole pitch period, not the closing phase.

To achieve high quality converted speech, excitation information needs to be taken into account. In this project, 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.

II. OVERVIEW OF PROPOSED SYSTEM

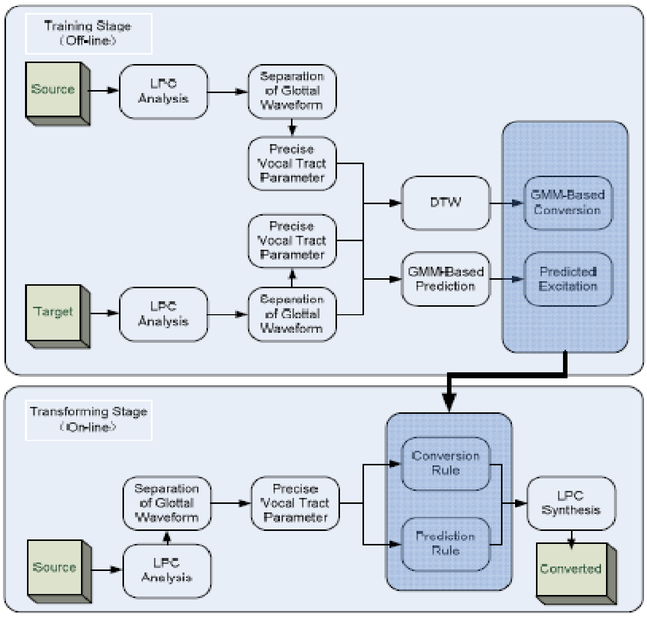
Training Stage (Off-line)
Source
LPC Analysis
Separation of Glottal Waveform
Precise Vocal Tract Parameter
DTW
GMM-Based Conversion
Precise Vocal Tract Parameter
GMM-Based Prediction
Predicted Excitation
Target
LPC Analysis
Separation of Glottal Waveform
Transforming Stage (On-line)
Separation of Glottal Waveform
Precise Vocal Tract Parameter
Conversion Rule
Prediction Rule
LPC Synthesis
Source
LPC Analysis
Converted

Fig1. Block diagram of 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, 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 vocal tract 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 [7]. In the transforming stage, 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.

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

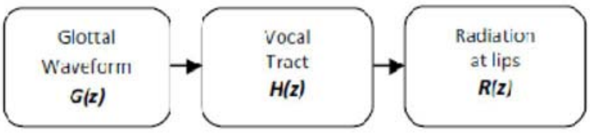

Fig.2. Speech production model

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 $R(z)$

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

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.,

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

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.

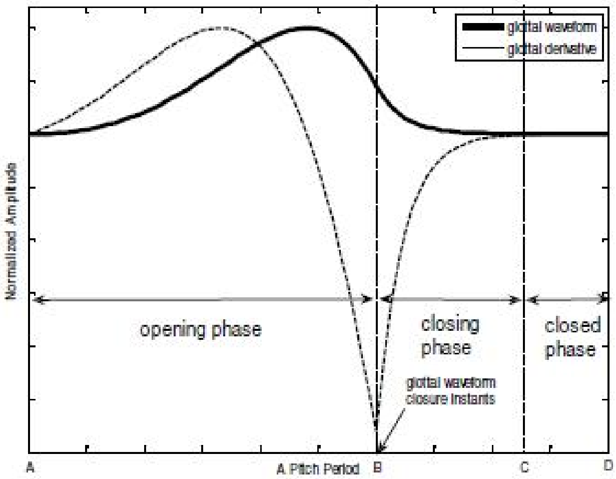

Fig.3. Glottal waveform and glottal derivative

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

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

Take this z-transform into time domain, we have

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

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

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

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,

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

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 analysed frames. Then a GMM model was used to fit the probability of the target speaker's vocal tract parameters as,

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

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*ω , 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 Mt denotes the target magnitude of the excitation signal of frame t, then the magnitude of the q th component of the model is,

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

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 **Mt** denotes the predicted excitation magnitude of the frame t.

IV. Performance Evaluation

Both objective and subjective experiments were performed to evaluate the performance of the proposed method. The speech corpus for this study consists of 20 sentences spoken by different male & Females in age limit of 20 to 50 years old which we refer to M and F respectively.

Training data: The 20 sentences are used as parallel speech data from both the source and the target speaker.

Tasks: 4 tasks; male to male, male to female, female to male and female to female.

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. Subjective Evaluation:

Hence in order to evaluate the proposed system on these two scales, there types of subjective measures are generally used.

- ABX test
- MOS test
- Similarity test

1) MOS & Similarity Test

Transformed speech is also generally evaluated in terms of naturalness and intelligibility by Mean Opinion Score (MOS)tests. In this test, the participants are asked to rank the transformed speech in terms of its quality and/or intelligibility. Listeners evaluate the speech quality of the converted voices using a 5-point scales, where:

- 5: Excellent
- 4:Good
- 3: Fair
- 2: Poor
- 1: Bad

This is similar to similarity test, but the major difference lies in the fact that we concentrate on the speaker characteristics in the similarity test & intelligibility in the MOS test.

2) ABX Text

ABX test is generally preferred over MOS & similarity tests we are more interested in accuracy than intelligibility or speakers characteristics. In order to check if the converted speech is perceived as the target speaker, ABX tests are most commonly used where participants listen to source(A), target (B) and transformed (x) utterances and are asked to determine whether A or B is closer to X in terms of speakers identity. A and B were either the target or the source speaker. Speakers A and B uttered the same sentence which, in general, was different from the sentence uttered by X. A score of 100% indicates that all listeners find the transformed speech closer to the target. For ABX test a latest version of Lacinato ABX/Shouter 2.35 is used. With ABX testing, target file (A) from a group of nearly-identical files & a target file (B) are secretly chosen and played and you try to determine which is same as that of morphed signal.

Table 1
Result of ABX test for traditional system

	Source	Target	Neither
Traditional	20.1%	33.6%	46.3%

The table gives the percentage of confidence & accuracy of guessing of the morphed signal. A 15 trials are taken where A & B are the source and target speaker respectively. And a morphed signal (C) is compared with source & target. Following results are obtained by guessing.

Table 2

Result of ABX test for proposed system

	confidence	Accuracy
Proposed	99.	99.

B) Objective Evaluation

Objective evaluations are indicative of conversion performance and useful to compare different algorithms within a particular framework. We have calculated the signal to noise ratio of the transformed speech.

1) 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

$$SNR = 10\log10\left(\frac{\sum s(t)^2}{\sum (s(t)-s_c(t))^2}\right) \qquad (11)$$

Where $s(t)$ and $s_c(t)$ denote target voice and converted voice separately. Table 1 shows the result of the comparison of the distortion of the traditional morphing system and the proposed system.

A Signal to Noise Ratio (SNR) of less than 1.1 suggests a very poor quality of signal. Using a proposed method we are getting a good quality signal having an SNR within range of 2.5 to 3.6.

Table 3

SNR of the systems

System	Traditional	Proposed
SNR(dB)	2.0471	2.9865

Hence in order to evaluate the proposed system, there types of objective measures are generally used.

V. 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.

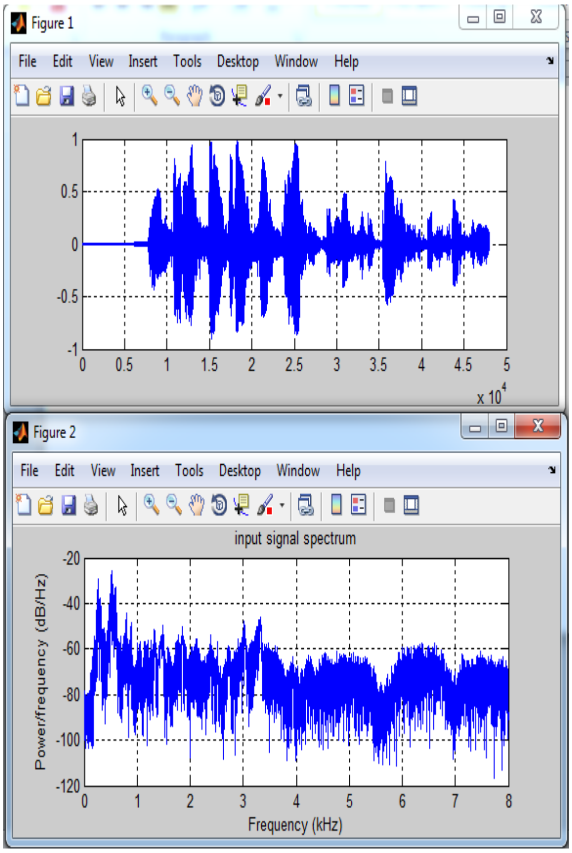

Fig 4. Input signal and its spectrum

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 4 and 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.

1) Time domain waveform comparison

The time domain of one frame of source and morphed signals is shown in figure 4 and 5. It is noticed that although the morphed signal has approximately the same time period as that of the target one but still there are noticeable difference in shape, duration & energy distribution.

The pitch periods of the signal source and morphed signal was monitored as illustrated in figure 4 and 5. Obviously pitch period of morphed signal is similar to that of target one.

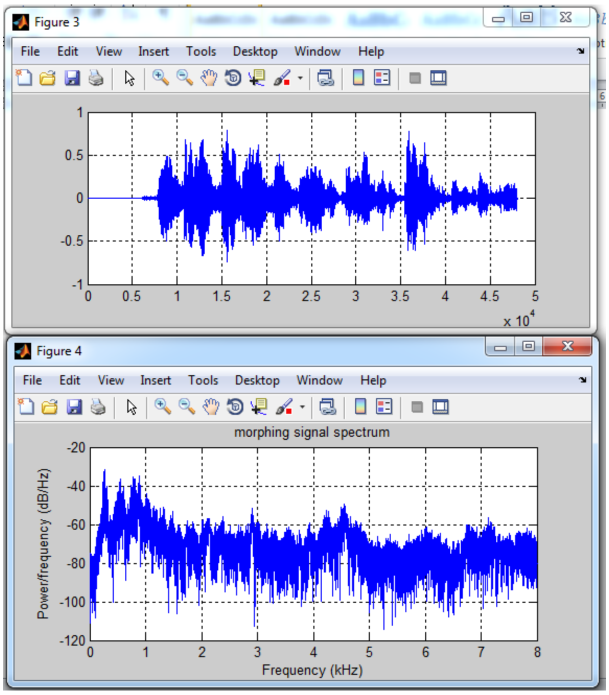

Fig5. Target Signal and Its spectrum

VI. Conclusions

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

I wish to express my sincere gratitude to Prof.S.L.Lahudkar, Electronics & Telecommunication Department, ICOER, Pune for his constant guidance throughout the course of the work and many useful discussions which enabled me to know the subtleties of the subject in proper.

I also thanks to the Principal Dr. S. V. Admane, ICOER, Pune, for providing me necessary facilities to carry out the work.

References

[1] be 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 noise model”, ICASSP, 1998, Seattle, Washington, USA , pp.281-284 ,

[5] Hui Ye and Steve Young, “High quality voice morphing”, ICASSP, 2004, Montreal, Canada [14] Nam I. H., “Voice personality transformation”, Ph.D. Thesis, Electrical Engineering, Rensselaer Polytechnic Institue ,Troy ,New York ,1991

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

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

[8] Duxans H., Bonafonte A., “Residual conversion versus prediction on voice morphing system”, ICASSP, 2006

[9] R. J. McAulay and T. F. Quatiery, “Sinusoidal coding,” in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., New York: Elsevier Science Inc., 1995.

[10]] K. Tanaka and M. Abe, “A new fundamental frequency modification algorithm with transformation of spectrum envelope according to f0,” in *Proc. ICASSP*, Munich, 1997, pp. 952-954

[11] L. M. Arslan, “Speaker transformation algorithm using segmental codebooks (STASC),” *Journal Speech Communication*, vol. 28, no. 3, pp. 211-226, Jul. 1999.

[12] [12] A. Mousa, “Voice conversion using pitch shifting algorithm by time stretching with PSOLA and re-sampling,” Journal of Electrical Engineering, vol. 61, no. 1, pp. 57-61, 2011.

[13] speech recognition,” International Journal for Advance Research in Engineering and Technology, vol. 1, no. VI, Jul. 2011.

[14] X. Lu and J. Dang, “An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification,” Speech Communication, vol. 50, no. 4, pp. 312-322, 2008.

[15] G. Xu, Q. Zou, D. Zhao, and D. Huang, “Straight model for voice conversion based on acoustical universal structure,” in Proc. International Conference on Audio, Language and Image Processing (ICALIP), Jul. 2012, pp. 454-458.

[16] X. Chen, W. Q. Zhang, and J. Liu, “An improved model for voice conversion based on Gaussian mixture model,” in Proc. International Conference on Computer Application and System Modeling (ICCASM), 2010.

[17] P. R. Gurumoorthy, “LPC based voice morphing,” University of Florida, 2006