HIGH QUALITY VOICE MORPHING

Hui Ye and Steve Young

Cambridge University Engineering Department
Trumpington Street, Cambridge, England, CB2 1PZ
hy216@eng.cam.ac.uk, sjy@eng.cam.ac.uk

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. Most of the recent approaches to voice morphing apply a linear transformation to the spectral envelope and pitch scaling to modify the prosody. Whilst these methods are effective, they also introduce artifacts arising from the effects of glottal coupling, phase incoherence, unnatural phase dispersion and the high spectral variance of unvoiced sounds. A practical voice morphing system must account for these if high audio quality is to be preserved.

This paper describes a complete voice morphing system and the enhancements needed for dealing with the various artifacts, including a novel method for synthesising natural phase dispersion. Each technique is assessed individually and the overall performance of the system evaluated using listening tests. Overall it is found that the enhancements significantly improve speaker identification scores and perceived audio quality.

1. INTRODUCTION

Voice morphing, also referred to as voice transformation and voice conversion, is a technique for modifying a source speaker's speech to sound as if it was spoken by some designated target speaker. There are basically three inter-dependent issues that must be addressed before building a voice morphing system. Firstly, a model is needed for decomposing and regenerating the speech signal. Secondly, the features of the model which encode speaker identity must be determined. Thirdly, the type of conversion function and the method of training and applying the conversion function must be decided.

Generally, voice morphing aims to control speaker identity independently of the message and the environment. This speaker identity is normally determined by the average pitch, the formant structure and the characteristics of the vocal tract. The vocal tract and formant characteristics can be represented by the overall shape of the spectral envelope and hence this is the key feature to transform in most voice conversion systems. Various approaches have been proposed for effecting the transformation including codebook mapping [1] and linear transformations. Of these, the linear transformation technique applied in conjunction with a sinusoidal speech model has been shown by Stylianou et al. [2] and Kain [3] to outperform other approaches in terms of speech quality.

Our system also uses a sinuisoidal model and an interpolated linear transformation approach[4]. However, evaluation of our baseline system showed that, even in pitch synchronous mode, simple transformations of the spectral envelope can introduce significant artifacts into the converted speech. Firstly, the averaging effect of the linear transformation suppresses spectral details resulting in muffled-sounding speech. To solve this problem, Kain used a residual prediction approach to regenerate the spectral detail in the converted spectral envelope[3] and here this approach is developed further. Spectral envelope distortion is not the only problem, however. The synthesis of natural phase dispersion and the transformation of unvoiced sounds also create problems which can degrade the quality of the converted speech.

In this paper, we identify and present solutions for each of the above problems, and then compare the performance of the enhanced system with the baseline. In section 2, the framework of our baseline system is briefly described, and then in section 3, the various problems and corresponding solutions are presented. The performance of the enhanced system with these new techniques integrated is then evaluated in section 4 before presenting overall conclusions in section 5.

2. BASELINE SYSTEM

Our voice morphing system uses a sinusoidal model for speech signal representation and modification. In principle, this model can support modifications to both the prosody and the spectral characteristics of the source signal without inducing significant artifacts[5], but in practice, conversion quality can be compromised by phase incoherency in the regenerated signal. Normally to avoid this problem, the pitch onset time for every speech frame has to be estimated [6]. However the accuracy of this estimation is rather low especially for weakly voiced sounds, and errors can result in significant distortion. In our system, this problem is avoided by using a pitch synchronous sinuisoidal model where each speech frame for modification is a single pitch period, and therefore can be regarded as an independent unit. The speech frame after modification can then be concatenated together without any phase incoherency problems. This of course leaves the residual problem of determining the pitch epochs in the source signal but we have found that this can be done more reliably than onset time estimation. Furthermore, in many applications, source data can be recorded with accompanying laryngograph waveforms.

Line spectral frequencies (LSF) are used to represent the spectral envelope. When compared to other features, such as cepstral coefficients [2] and discrete line spectrum using cubic spline interpolation [4], LSF requires less coefficients to efficiently capture the formant structure and it has better interpolation properties.

The VOICES database from OGI is used for evaluation[3]. This corpus contains recorded speech from 12 different speakers reading 50 phonetically rich sentences. Each sentence is spoken 3 times by each speaker. The recording procedure involved a "mimicking" approach which resulted in a high degree of natural

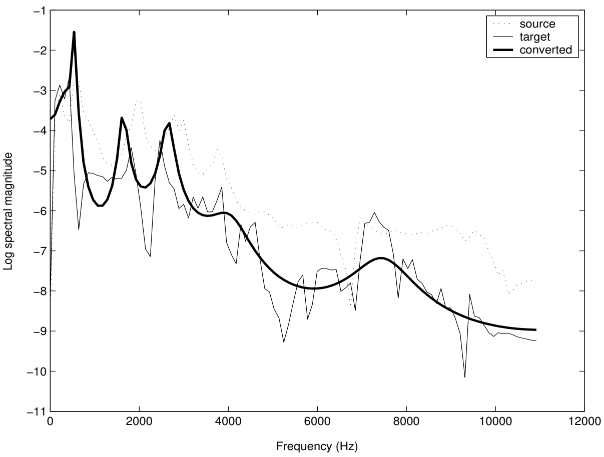

Fig. 1. *Spectral envelope conversion.*

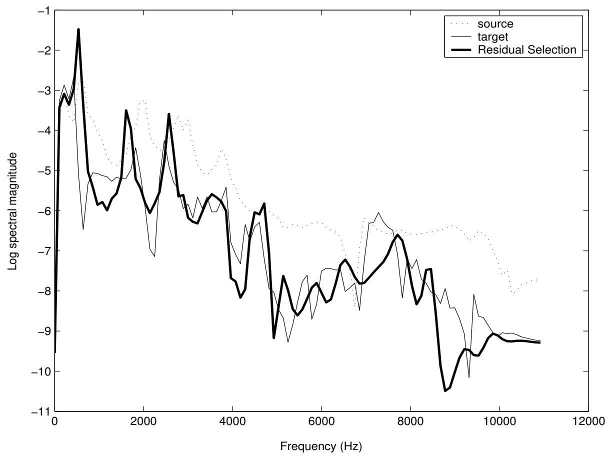

Fig. 2. *Spectral envelope conversion using residual selection.*

time-alignment between different speakers. Pitch period information for each utterance is also provided and this was used for our pitch synchronous speech representation. In our experiments, four different voice conversion tasks were investigated: male-to-male, male-to-female, female-to-male and female-to-female conversion. For each task, we used the first 120 utterances as training data, and the remaining 30 utterances as the test set. A DTW algorithm was used to align the corresponding utterances before training and testing.

3. SYSTEM ENHANCEMENT

The converted speech produced by the baseline system described above will often contain artifacts. This section discusses these in more detail and describes the solutions developed to mitigate them.

3.1. Residual Selection

Fig.1 shows an example of envelope conversion using the linear transformation approach. Although the formant structure of the source speech has been transformed to more closely match the target, the spectral detail has been lost as a result of reducing the dimensionality of the envelope representation during the transform. Another clear effect is the broadening of the spectral peaks caused, at least in part, by the averaging effect of least square error estimation. All these degradations lead to muffled effects in the converted speech.

To solve this problem, a straightforward idea is to reintroduce the lost spectral details to the converted envelopes. Our method for doing this is called Residual Selection in which residuals are selected from a database extracted from the training data. This method is a refinement of the residual codebook method proposed in [3].

The log magnitude spectrum of the spectral residual r_t is calculated via

$$r_t = 20log_{10}H(t)_{sin} - 20log_{10}H(t)_{env} \tag{1}$$

where $H(t)_{sin}$ is the amplitude contour of the sinusoidal components of speech frame t and $H(t)_{env}$ is the spectral envelope represented by the LSF coefficients. Each spectral residual r_t is associated with a vector $v_t = [f_1, f_2, \cdots, f_d, \Delta f_1, \Delta f_2, \cdots, \Delta f_d]'$, where f_i are the line spectral frequencies, and Δf_i are the differences between speech frames $t-1$ and t. The r_t and v_t corresponding to all speech frames in the training data are then gathered together to form a database.

The criteria used to select the appropriate residual r_k for each converted spectral envelope is to choose that residual whose associated v_k minimizes the following square error

$$\mathcal{E} = (v_k - \tilde{v})'.(v_k - \tilde{v}) \tag{2}$$

where $\tilde{v}$ is the spectral vector associated with the converted spectral envelope.

Using residual selection, the spectral distortion on the OGI test set was reduced from -5.6dB to -7.2dB using this method (0dB corresponds to the initial distortion between the unconverted source and target envelopes). Fig.2 shows an example of envelope conversion using this approach where it can be seen that some measure of spectral detail has been successfully reintroduced.

3.2. Phase Prediction

The spectral magnitudes and phases of human speech are highly correlated, and when only spectral magnitudes are modified and the original phases preserved, a harsh quality is introduced into the converted speech. However, to simultaneously model the magnitudes and phases and then convert them both via a single unified transform is extremely difficult. Since phase dispersion actually determines waveform shape, if we can predict the waveform shape based on the spectral envelope then we can also predict the phases. Inspired by this idea, the following phase prediction approach has been developed.

A GMM model is first trained to cluster the target spectral envelopes coded via LSF coefficients into M classes $(C_1, \cdots, C_M)$. Then for each target envelope v we have a set of posterior probabilities

$$P(C_m|v) = \frac{\alpha_m N(v; \mu_m, \Sigma_m)}{\sum_{i=1}^{M} \alpha_i N(v; \mu_i, \Sigma_i)}, m = 1, \cdots, M \tag{3}$$

where $\{\alpha_i\}$, $\{\mu_i\}$ and $\{\Sigma_i\}$ are the mixture weights, mean vectors and covariance matrices of the GMM model respectively.The vector $\mathcal{P}(v)$ composed from these probablities can then be regarded

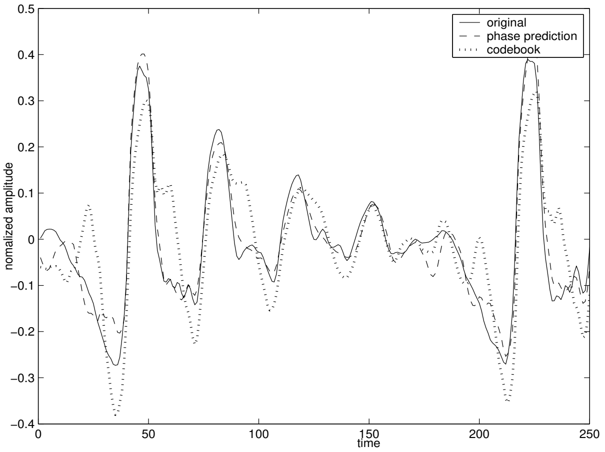

Fig. 3. *A small segment of an original signal, the output of phase prediction and the codebook quantization.*

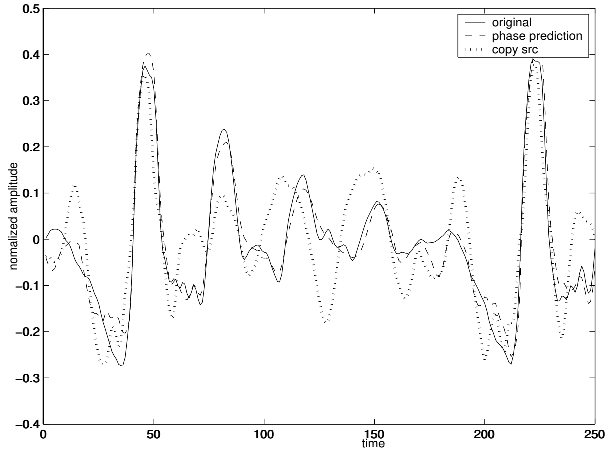

Fig. 4. *A small segment of an original target signal, the output of phase prediction and the output of copying phases from source.*

as another form of representation of the spectral shape,

$$\mathcal{P}(v) = [P(C_1|v), \cdots, P(C_M|v)]' \tag{4}$$

Each element $P(C_i|v)$ of this vector can be regarded as the weight of a codebook entry T_i and the set of M codebook entries

$$\mathcal{T} = [T_1, \cdots, T_M] \tag{5}$$

can be chosen to minimise the coding error over the training data. That is, $\mathcal{T}$ can be chosen to minimize the following least square error criterion,

$$E = \sum_{t=1}^{N} (s(t) - \mathcal{T}\mathcal{P}(v_t))'(s(t) - \mathcal{T}\mathcal{P}(v_t)) \tag{6}$$

where $s(t)$ is the t'th speech frame in the target training data normalized to a certain pitch value, say 100Hz. The standard solution to equation (6) is then

$$\mathcal{T} = \Big(\sum_{t=1}^{N} s(t)\mathcal{P}(v_t)'\Big)\Big(\sum_{t=1}^{N} \mathcal{P}(v_t)\mathcal{P}(v_t)'\Big)^{-1} \tag{7}$$

Having estimated $\mathcal{T}$ from the training data, the waveform shape of any converted spectral envelope can be predicted as

$$\tilde{s}(t) = \mathcal{T}\mathcal{P}(\tilde{v}_t); \tag{8}$$

The required phases can then be obtained from the predicted waveform $\tilde{s}(t)$ using the analysis routine and pitch-scale modification algorithm of sinusoidal modelling.

Table 1 shows the signal to noise ratio (SNR) using three different phase coding methods: copying the phase spectra of the source speech, using a phase codebook [8] and the new phase prediction approach. The latter clearly outperforms the other two approaches and furthermore the improvement in audio quality is apparent in listening tests. Examples of these methods when applied to a short segment of speech are shown in Fig. 3 and Fig. 4.

Table 1. *The SNR ratio in dB of different phase coding methods.*

src phases	codebook phases	phase prediction
3.2171	6.1544	7.2079

3.3. Transforming Unvoiced Sounds

Unvoiced sounds contain very little vocal tract information and their inclusion in the envelope transformation process results in noticeable degradation. Hence, in common with other transform-based systems, unvoiced sounds in the baseline system are simply copied to the target. Many unvoiced sounds do, however, have some vocal tract colouring and simply copying the source to the target affects the converted speech characteristics, especially in cross gender conversion. A typical effect is the perception of another speaker whispering behind the target speaker.

Since most unvoiced sounds have no obvious vocal tract structure and cannot be regarded as short term stationary signals, their spectral envelopes show large variations, therefore it is not effective to convert them using the same solution as for the voiced sounds. However it can be shown empirically that randomly deleting, replicating and concatenating segments of the same unvoiced sound does not induce significant artifacts. This observation suggests a possible solution based on unit selection and concatenation to transform unvoiced sounds.

In this approach, a GMM model is first trained on the LSF-encoded spectral envelopes of the unvoiced target speech frames and then used to label all the unvoiced frames. The unvoiced speech frames associated with each GMM label are then gathered together into a database.

When a segment of n successive unvoiced speech frames from the source speaker needs to be transformed, these n frames are first labelled using the GMM model. According to the labels, target unvoiced frames are then chosen from the database using a criterion that encourages the selection of frames which were adjacent in the original target data. This is done by successively selecting the longest matching model sequence. For example, if the sequence of source labels is "1 1 1 3 3 2 1" , and the longest matching sequence in the target database is "1 1 1 3" then the speech frames corresponding to this subsequence are extracted. The procedure then repeats looking for a match for "3 2 1" and so on until the whole of the source segment is matched. The extracted target frames are then concatenated and their amplitudes are modified to match the original source frames.

This method of transforming unvoiced sounds eliminates the whispering artifacts in the converted speech, however there are still some discontinuities which require further attention.

3.4. Post-filtering

As mentioned earlier, transform-based voice conversion systems have a tendency to broaden the formants in the target speech. To mitigate this effect and suppress noise in the spectral valleys, a final post-processing stage applies a perceptual filter to the regenerated spectral envelope of all voiced sounds:

$$H(\omega) = \frac{A(z/\beta)}{A(z/\gamma)}, 0 < \gamma < \beta \leq 1 \tag{9}$$

where $A(z)$ is the LPC filter and the choice of parameters in our system is $\beta = 1.0$ and $\gamma = 0.94$. This filter is popular in speech coding [9] and its more general use in voice conversion is discussed in [4].

4. EVALUATION

In section 3, the individual enhancements to the baseline system were tested using a variety of objective measures and in each case, an improvement was indicated. In order to test the overall subjective quality, listening tests were conducted to assess both the perceptual accuracy of the transformation, i.e. does the transformed source sound like the target speaker, and the audio quality.

For the former, an ABX-style preference test was performed whereby a panel of 23 listeners were asked to judge whether an utterance X sounded closer to utterance A or B in terms of speaker identity, where X was the converted speech and A and B were either the source speech or the target speech. The source and target were chosen randomly from both male and female speakers. There were 32 transformed utterances in total, equally split between within-gender and cross-gender transformations. Table 2 gives the percentage of the converted utterances that were labelled as closer to the target for each case, where the “enhanced system” refers to the enhanced system that integrates all the above techniques. The results clearly show that the new system outperforms the baseline system in terms of transforming the speaker identity. This is probably mostly due to the inclusion of the spectral residual which contains speaker specific information. It is also interesting but perhaps not surprising to note that almost all the errors occurred in the within-gender transformations.

To assess speech quality between the baseline system and the new system, a second preference test was conducted whereby listeners were presented each time with a pair of utterances generated by the baseline system and the new system respectively, and then listeners were asked to judge which one has the better speech quality. Table 3 indicates that most listeners are in favor of the converted speech generated by the enhanced system. This is consistent with the previous objective evaluations.

Table 2. *Results from the ABX test.*

	baseline system	enhanced system
ABX	86.4%	91.8%

Table 3. *Results from the preference test.*

	baseline system	enhanced system
preference	38.9%	61.1%

5. CONCLUSION

This paper has presented a complete solution for high quality voice morphing based on a limited amount of parallel speech data. Building on the well-established approach of using interpolated linear transforms to convert the spectral envelope, the importance of three other factors has been highlighted: residual selection, phase prediction and the conversion of unvoiced sounds. In each case, effective solutions have been presented and assessed objectively and subjectively. Both the fidelity and the quality of the enhanced system has been shown to be significantly better than the baseline system demonstrating the need for good solutions to these issues. Overall the quality is judged to be sufficient for commercial applications which require medium fidelity such as those which operate over the telephone. However, there appears to be still some way to go before these techniques can support high fidelity studio applications.

6. ACKNOWLEDGMENTS

This work was supported by a grant from Anthropics Technology Ltd. The authors thank the volunteers of the perceptual tests for their assistance.

7. REFERENCES

[1] Abe, M., Nakamura, S., Shikano, K. and Kuwabara, H., “Voice conversion through vector quantization”, Proc. IEEE ICASSP, 1988.

[2] Stylianou, Y., Cappe, O. and Moulines, E., “Continuous probabilistic transform for voice conversion”, IEEE Trans. on Speech and Audio Processing, vol. 6, no. 2, pp. 131-142, 1998.

[3] Kain, A.. “High resolution voice transformation”, PhD dissertation, OGI, 2001.

[4] Ye, H. and Young, S.. “Perceptually Weighted Linear Transformation for Voice Conversion”, Eurospeech 2003.

[5] Quatieri, T.F. and McAulay, R.J., “Shape invariant time-scale and pitch modification of speech”, IEEE Transactions on Acoustics, Speech and Signal Processing, pp. 497-510, 1992.

[6] Macon, M.W., “Speech Synthesis Based on Sinusoidal Modeling” , PhD thesis, Georgia Institute of Technology, October 1996.

[7] McAulay, R.J. and Quatieri, T.F., “Sine-Wave Phase Coding at Low Data Rates”, Proc. IEEE ICASSP, vol. 1, pp. 577-580, 1991.

[8] McAulay, R.J. and Quatieri, T.F., “Phase Modelling and Its Application to Sinusoidal Transform Coding”, Proc. IEEE ICASSP, pp. 1713-1715, 1986.

[9] Chen, J.H. and Gersho, A., “Real-time vector APC speech coding at 48000 bps with adaptive postfiltering”, in Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, 1987.