Voice Morphing Using the Generative Topographic Mapping

Christina ORPHANIDOU
Oxford Centre for Industrial and Applied Mathematics
Mathematical Institute, University of Oxford
Oxford OX1 3LB, U.K.

and

Irene M. MOROZ
Oxford Centre for Industrial and Applied Mathematics
Mathematical Institute, University of Oxford
Oxford OX1 3LB, U.K.

and

Stephen J. ROBERTS
Robotics Research Group
Department of Engineering Science, University of Oxford
Oxford OX1 3PJ, U.K.

ABSTRACT

In this paper we address the problem of Voice Morphing. We attempt to transform the spectral characteristics of a source speaker's speech signal so that the listener would believe that the speech was uttered by a target speaker. The voice morphing system transforms the spectral envelope as represented by a *Linear Prediction* model. The transformation is achieved by codebook mapping using the *Generative Topographic Mapping*, a non-linear, latent variable, parametrically constrained, Gaussian Mixture Model.

Keywords: Voice Morphing, LPC, GTM, codebook mapping.

1. INTRODUCTION

Voice morphing technology enables a user to transform one person's speech pattern into a different pattern with distinct characteristics while preserving the original meaning. The new characteristics are, in most applications, those of another speaker. Voice morphing technology has numerous applications such as text-to-speech adaptation, where the voice morphing system can be trained on relatively small amounts of data and allow new voices to be created at a much lower cost than the currently existing systems. The voice morphing system can also be used in the situation where the speaker was not available and previous recordings had to be used. Other applications include voice disguise as well as low bandwidth speech encoding, where speech may be transmitted without revealing the speaker's identity and then re-synthesised.

The standard approach to the problem usually comprises a training phase where some training speech data from the source and target speakers are used in order to create a transformation that maps the speech space of the source speaker to that of the target speaker. The most popular features used for voice conversion are the formant frequencies [1]. These are essentially the characteristic resonant frequencies of the vocal tract [2]. The transformation is in general based on *codebook mapping*[1,3], i.e., a one to one correspondence between the codebook entries of the source and target speakers. Possible problems that these methods might encounter are artefacts introduced at the boundaries between speech frames, limitations on the robust estimation of parameters or deformation introduced during synthesis of the target speech. Furthermore, transformation of the glottal excitation as well as the vocal tract characteristics has been the subject of several recent studies [3,4,5].
However, since it is generally assumed that the vocal tract characteristics are the major features controlling speaker identity, in this study we propose a new and effective solution to converting the formant characteristics with the goal of maintaining high speech quality.

2. DESCRIPTION OF THE SYSTEM

Our voice morphing system consists of three stages: the training stage, the conversion stage and the synthesis stage. The flow diagram of the proposed voice morphing system is shown in Figure 1.

In the training stage, speech signals of the source and target speakers are analysed and the voice characteristics are extracted by means of a mathematical optimisation technique that is very popular in the speech processing world, the Linear Prediction Coding (LPC) technique [6]. Linear Prediction Coding models the speech signal as a linear combination of past values of a hypothetical input to a system whose output is the given signal. By minimising the sum of the squared differences between the actual speech samples and the linearly predicted ones, a unique set of predictor coefficients are determined. These high-dimensional predictor coefficients define the acoustic space for each speaker and relate closely to formant frequencies. Furthermore, the LPC residual can be a reasonable approximation to the glottal excitation signal.
In order to increase the robustness of the conversion, the high dimensional speech spaces, where the voice characteristics lie, will be described by a continuous probability density

corresponding to a parametric constrained Gaussian Mixture Model that is topologically orientated by means of the Generative Topographic Mapping (GTM)[7]. The underlying idea of the GTM is that the high-dimensional data variables we observe are being generated from a smaller number of hidden or latent variables. The model allows for the relationship between the latent and the data space to be non-linear by considering a non-linear, parameterised mapping from the latent space to a corresponding point in the data space.

In the conversion stage, the acoustic features of the sound signal as represented by the GTM, are transformed to those of the target speaker by means of codebook mapping. This is a one-to-one correspondence between the codebook entries of the source speaker and the target speaker. The transformed codebook then forms a different linear prediction filter.

Finally, in the synthesis stage, the new linear prediction filter excites the glottal excitation (the LPC residual) of the source speaker in order to produce the desired speech.

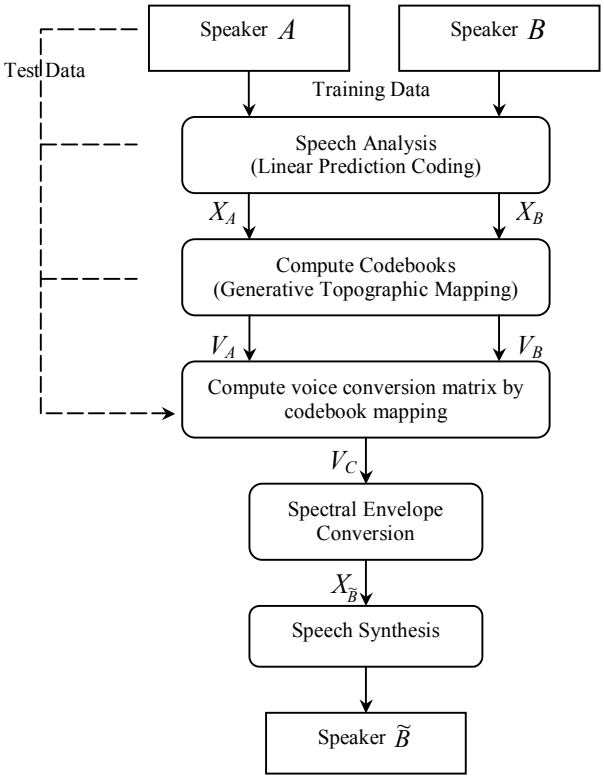

Figure 1: Flow diagram of our voice morphing system

3. TRAINING

Linear Prediction Analysis

The reasons for selecting LPC analysis are the similarity between the source/filter decomposition, yielded by the mathematics of linear prediction and the source/filter model of the human vocal tract; the fact that it can encode good quality speech at a low bit rate ; and the high accuracy of its estimates.

The training speech (sampled at 10kHz) from source and target speakers was first segmented into phonemes and time-aligned.

12^{th} order LPC analysis was performed on a frame-by-frame basis in order to represent spectral characteristics of source and target speaker vocal tract characteristics.

LPC analysis was performed using VOICEBOX [9].

Figure 2 shows the LPC parameters for a vowel /a/ from a British female speaker. The phoneme was broken into 8 frames and 12 LPC coefficients were calculated for each frame

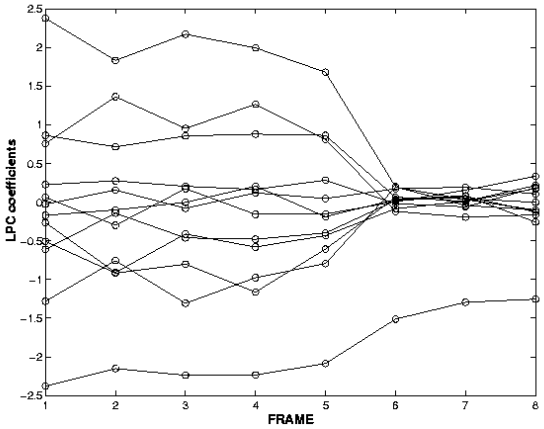

Figure 2: 12^{th} order LPC parameters for a vowel /a/ segmented in 8 frames from a British female speaker.

Next, the linear prediction coefficients were modelled by the Generative Topographic Mapping.

Generative Topographic Mapping (GTM)

The following is based on [7].

The GTM defines a non-linear, parametric mapping $y(x,W)$ from an L-dimensional latent space $x \in \Re^L$ to a D-dimensional data space $y \in \Re^D$ where normally $L < D$. The mapping $y(x,W)$ that maps every point in latent space to a point in the data space is defined as continuous and differentiable. These data will be confined to an L-dimensional non-Euclidean manifold embedded within the D-dimensional data space. If we define a probability distribution $p(x)$ on the latent-variable space, a corresponding distribution $p(y \mid W)$ in the data space can be convolved with a isotropic Guassian noise distribution :

$$p(t \mid x, W, \beta) = \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left\{-\frac{\beta}{2}\|y(x,W) - t\|^2\right\} \qquad (1)$$

where t is a point in the data space and β^{-1} denotes the noise variance.

The t-distribution in data space, for given parameters β and W, is then obtained by integrating over the x-distribution in latent space

$$p(t \mid W, \beta) = \int p(t \mid x, W, \beta) p(x) dx \qquad (2)$$

For a set of independent identicallly distributed data points $t_1, \cdots, t_N$, we can determine the parameter matrix W and the inverse variance β using maximum likelihood by the following log-likelihood function:

$$L = \prod_{n=1}^{N} p(t_n \mid W, \beta) = \prod_{n=1}^{N} \left[\frac{1}{K} \sum_{k=1}^{K} p(t_n \mid x_k, W, \beta) \right] . \quad (3)$$

Because the integral of (3) is difficult to be solved analytically, we here consider a specific form for $p(x)$ given by a sum of delta functions centred on the nodes of a regular grid in latent space

$$p(x) = \frac{1}{K} \sum_{k=1}^{K} \delta(x - x_k) \quad (4)$$

So, we can rewrite Eq(4) and Eq(5) as

$$p(t \mid W, \beta) = \frac{1}{K} \sum_{k=1}^{K} p(t \mid x_k, W, \beta) \quad (5)$$

and

$$\ell = \sum_{n=1}^{N} \ln \left(\left[\frac{1}{K} \sum_{k=1}^{K} p(t_n \mid x_k, W, \beta) \right] \right). \quad (6)$$

In order to maximise this function we employ the Expectation-Maximisation (EM) algorithm that computes the expectations of sufficient statistics of the latent variables for fitting the model to the data variables.

The underlying idea of the GTM is that the high-dimensional data variables that we observe are generated from a smaller number of hidden or latent variables. The GTM allows for the relationship between the latent and the data spaces to be non-linear by considering a non-linear parameterised mapping from the latent space to a corresponding point in the data space. Assuming that this mapping is smooth, these points will be confined to a low dimensional manifold in the high-dimensional data space. By then defining a distribution over the latent space, a corresponding distribution over the manifold in the data space will be induced, establishing a probabilistic relationship between the two spaces. After the model is fitted, the points in the data space will be related to points on the curved manifold that correspond to points in the latent space. This way, a low-dimensional representation of the points in the high dimensional data space can be established.

Application of the GTM to the LPC coefficients

The LPC coefficients were fitted to the GTM model based on [7] and using the GTM Toolbox [10].

Implementation took place as follows:

1) A square grid of latent points was generated in the 2-D latent space centred at the origin in the latent space.
2) A grid of basis function centres was generated also centred at the origin in the latent space.
3) The basis function width was chosen.
4) The matrix of basis function activations was calculated.
5) The weight matrix was initialised using Principal Component Analysis.
6) The inverse variance was initialised.
7) The algorithm was trained for 40 iterations.

After 40 iterations our latent variable sample consisted of a set of latent points, each of which corresponded to one of the centres of Gaussian mixture that were generated by the GTM and lie on a 2-D manifold in the 12-D speech space.
These vectors are the *codebook vectors* of each speaker's speech space for each phoneme.

4. CONVERSION

Formation of the transformed codebooks

For each test phoneme, the posterior probability distribution of each frame was calculated over the latent sample and the closest codebook vector from the source codebook was determined. The transformed codebooks were then created by means of *codebook mapping*. The transformation matrix was formed by replacing the closest codebook vector of each frame by the corresponding codebook vector of the target speaker. Essentially, we substituted the vector space of the source speech with that of the target speech, and moved all the vectors in the source vector space to the desired points in the target vector space.

Figure 3 shows the posterior probability distribution over the latent space for a given frame. For the given frame, the closest codebook vector is obviously the 7th one. It is therefore replaced by the corresponding one in the target space.

The transformed codebooks are then converted to prediction coefficients, which will then filter the excitation signal in order to synthesise speech.

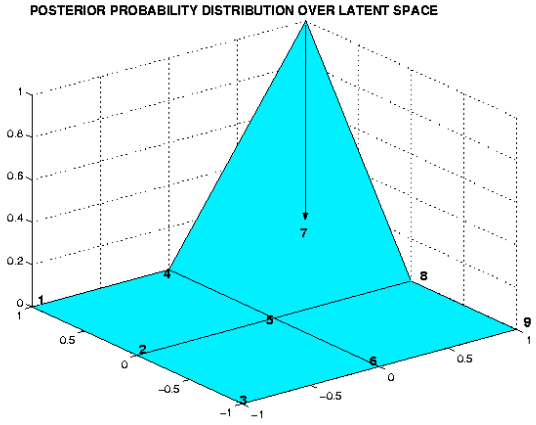

Figure 3: Determination of closest codebook vector to test data vectors. The posterior probability distribution of the data point determines the closest latent point and its corresponding codebook vector.

5. SPEECH SYNTHESIS

The morphed speech was synthesised from the linear prediction parameters. The system used had the same parametric representation as that used in extracting the features of the original speech signals. The control parameters supplied to the synthesiser were the *excitation signal* in the form of impulses and the morphed LPC coefficients. Since the speech signal was analysed frame-by-frame it was also synthesised frame-by-frame. The speech synthesis technique we used was essentially a filtering technique where the morphed LPC coefficients act as a filter to the excitation signal and the output of this is the synthesised speech signal.

Determination of glottal excitation signal

The *LPC residual* can be a reasonable approximation to the glottal excitation signal [3]. Based on the prediction coefficients, an inverse filter, A(z), is formulated as:

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \quad . \qquad (7)$$

This filter was used to estimate an approximation to the excitation signal for the source speaker.

6. RESULTS AND EVALUATION

The samples tested consisted of two voiced phonemes (/a/ and /o/) taken from the word "ago". Samples consisted of one British female, one American female and one British male.

Figure 4 shows the source, target and morphed signals of phoneme /o/ with the British and American females as the source and target speakers respectively. The LPC analysis was 12^{th} order and the frame size 50 samples. A grid of 9 latent points was used with 4 basis functions.

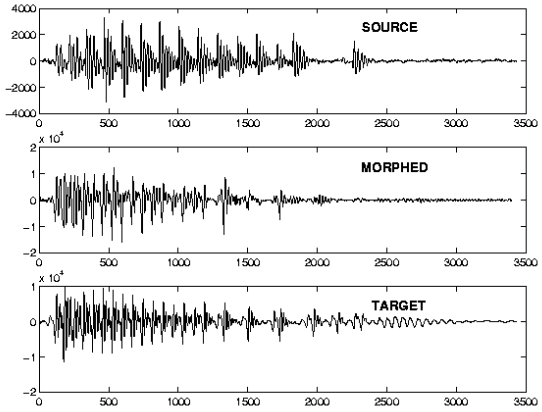

Figure 4: Converted results for phoneme /o/ between the two female speakers. At the top, the source signal is presented and at the bottom the target signal. The morphed signal is between these two signals.

Subjective Evaluation

The original and modified speech signals were presented to a number of independent listeners. They were asked to identify the target speaker and to rate the perceptual quality on a five-point scale. The listeners identified the target speakers with a 90% success rate and the average quality score was 3.6 for the British female-American female transformation and 3.2 for the British female-British male transformation.

Parameter optimisation

We tested our proposed method by calculating the mean square errors with different GTM parameters (number of latent points and basis functions). Figure 5 depicts the normalised mean square errors of the morphed speech in comparison to the target speech for the two different phonemes. For small test samples it is evident that a smaller number of latent points and basis functions gives better results due to the fact that they prescribe a smoother mapping.

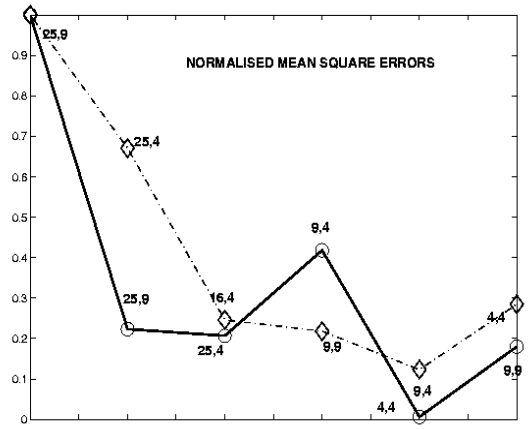

Figure 5: normalised mean square errors for different GTM parameter values for phoneme /a/ (dashed line) and phoneme /o/ (solid line).

7. CONCLUSION

In this study, a new voice conversion system has been presented. The system is based on codebook mapping of Linear Prediction coefficients modelled by the Generative Topographic Mapping. Perceptual tests confirmed that the spectral conversion was perceived as successful, since we managed to produce speech recognised as the target speaker. The performance of the method can be improved by also converting the residual waveform, which is an approximation to the glottal excitation. Also, using an increased training sample promises improvement of the speech listening quality.

8. REFERENCES

[1] M. Abe, S. Nakamura,K.Shikano and H. Kuwabara, "Voice Conversion through Vector Quantization", **Proc. IEEE ICASSP**, 1988, pp 565-568.

[2] L.R. Rabiner, R.W. Schafer, **Digital Processing of Speech Signals**, Prentice-Hall Signal Processing Series, 1978.

[3] L.M. Arslan, D.Talkin, "Voice conversion by codebook mapping of line spectral frequences and excitation spectrum", **Proceedings EUROSPEECH**, 1997,3:1347-1350.

[4] L.M. Arslan, "Speaker transformation algorithm using segmental codebooks (STASC)", **Speech Communication**, 1999, 28:211-226.

[5] K.S.Lee, D.H. Youn, I.W. Cha, "A new voice transformation method based on both linear and nonliear prediction analysis", **Proceedings ICSLP**, 1996, 1401-1404.

[6] J. Makhoul, "Linear Prediction: A tutorial review", **Proceedings IEEE**, 1975,Vol. 64, No. 4, pp. 561-580.

[7] J. Svensen, **GTM: The generative topographic mapping,** PhD thesis, Aston University, UK, April 1998.

[8] O. Turk, L.M Arslan, "Subband based voice conversion**" ICSLP 2002**, Denver, Colorado, USA.

[9] Mike Brookes, **VOICEBOX: Speech Processing Toolbox for MATLAB** (software).

[10] Markus Svensen , **The GTM Toolbox**, 1996 (software).