Quality-enhanced Voice Morphing using Maximum Likelihood Transformations

Hui Ye, *Student Member, IEEE,* and Steve Young, *Member, IEEE*

***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. The core process in a voice morphing system is the transformation of the spectral envelope of the source speaker to match that of the target speaker and linear transformations estimated from time-aligned parallel training data are commonly used to achieve this. However, the naive application of envelope transformation combined with the necessary pitch and duration modifications will result in noticeable artifacts. This paper studies the linear transformation approach to voice morphing and investigates these two specific issues. Firstly, a general maximum likelihood framework is proposed for transform estimation which avoids the need for parallel training data inherent in conventional least mean square approaches. Secondly, the main causes of artifacts are identified as being due to glottal coupling, unnatural phase dispersion and the high spectral variance of unvoiced sounds, and compensation techniques are developed to mitigate these. The resulting voice morphing system is evaluated using both subjective and objective measures. These tests show that the proposed approaches are capable of effectively transforming speaker identity whilst maintaining high quality. Furthermore, they do not require carefully prepared parallel training data.**

***Index Terms*— Voice morphing, voice conversion, linear transformation, phase dispersion**

I. Introduction

Voice morphing which is 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 many applications of voice morphing including customising voices for TTS systems, transforming voice-overs in adverts and films to sound like that of a well-known celebrity, and enhancing the speech of impaired speakers such as laryngectomees. Two key requirements of many of these applications are that firstly they should not rely on large amounts of parallel training data where both speakers recite identical texts, and secondly, the high audio quality of the source should be preserved in the transformed speech.

The core process in a voice morphing system is the transformation of the spectral envelope of the source speaker to match that of the target speaker and various approaches have been proposed for doing this such as codebook mapping [1], [2], formant mapping [3] and linear transformations [4], [5], [6]. Codebook mapping, however, typically leads to discontinuities in the transformed speech. Although some discontinuities can be resolved by some form of interpolation technique [2], the conversion approach can still suffer from a lack of robustness as well as degraded quality. On the other hand, formant mapping is prone to formant tracking errors. Hence, transformation-based approaches are now the most popular. In particular, the continuous probabilistic transformation approach introduced by Stylianou et al. [4] provides the baseline for modern systems. In this approach, a Gaussian mixture model (GMM) is used to classify each incoming speech frame, and a set of linear transformations weighted by the continuous GMM probabilities are applied to give a smoothly varying target output. The linear transformations are typically estimated from time-aligned parallel training data using least mean squares. More recently, Kain has proposed a variant of this method in which the GMM classification is based on a joint density model[5]. However, like the original Stylianou approach, it still relies on parallel training data. Although the requirement for parallel training data is often acceptable, there are applications which require voice transformation for non-parallel training data. Examples can be found in the entertainment and media industries where recordings of unknown speakers need to be transformed to sound like well-known personalities. Further uses are envisaged in applications where the provision of parallel data is impossible such as when the source and target speaker speak different languages.

This paper begins by expressing the continuous probabilistic transform of Stylianou as a simple interpolated linear transform. Expressed in a compact form, this representation then leads straightforwardly to the realisation of the conventional training and conversion algorithms. In analogy to the transform-based adaptation methods used in recognition [7], [8], the estimation of the interpolated transform is then extended to a maximum likelihood formulation which does not require that the source and training data be parallel.

Although interpolated linear transforms are effective in transforming speaker identity, the direct transformation of successive source speech frames to yield the required target speech will result in a number artifacts. The reasons for this are as follows. Firstly, the reduced dimensionality of the spectral vector used to represent the spectral envelope and the averaging effect of the linear transformation result in formant broadening and a loss of spectral detail. Secondly, unnatural phase dispersion in the target speech can lead to audible artifacts and this effect is aggravated when pitch and duration are modified. Thirdly, unvoiced sounds have very high variance and are typically not transformed. However, in that case, residual voicing from the source is carried over to the target speech resulting in a disconcerting background whispering effect.

To achieve high quality of voice conversion, all these issues have to be taken into account and in this paper, we identify and present solutions for each of them. These include a spectral

refinement approach to compensate the spectral distortion, a phase prediction method for natural phase coupling and an unvoiced sounds transformation scheme. Each of these techniques is assessed individually and the overall performance of the complete solution evaluated using listening tests. Overall it is found that the enhancements significantly improve speaker identification scores and perceived audio quality.

The remainder of the paper is organised as follows. First, the transform-based voice morphing framework is outlined in Section II, followed by a description of the interpolated linear transform and its estimation under different training conditions. In Section III, the various problems discussed above and their corresponding solutions are presented. The performance of the enhanced system with these new techniques integrated is evaluated in Section IV and finally, overall conclusions are presented in Section V.

II. Transform Based Voice Morphing System

A. Overall Framework

Transform-based voice morphing technology converts the speaker identity by modifying the parameters of an acoustic representation of the speech signal. It normally includes two parts, the training procedure and the transformation procedure. The training procedure operates on examples of speech from the source and the target speakers. The input speech examples are first analyzed to extract the spectral parameters that represent the speaker identity. Usually these parameters encode the short-term acoustic features, such as the spectrum shape and the formant structure. After the feature extraction, a conversion function is trained to capture the relationship between the source parameters and the corresponding target parameters. In the transformation procedure, the new spectral parameters are obtained by applying the trained conversion functions to the source parameters. Finally, the morphed speech is synthesized from the converted parameters.

Although it is outside the scope of this paper, mapping the prosody of the source speaker to be like the target speaker is an equally important and challenging problem. In all of the work reported in this paper, the source pitch is simply shifted and scaled to match the mean and variance of the target speaker. This is just about adequate for similar speakers such as those used in the evaluations reported later in the paper but it is clearly not a general solution.

There are three inter-dependent issues that must be decided before building a voice morphing system. Firstly, a mathematical model must be chosen which allows the speech signal to be manipulated and regenerated with minimum distortion. Previous research [9], [4], [5] suggests that the sinusoidal model is a good candidate since, in principle at least, this model can support modifications to both the prosody and the spectral characteristics of the source signal without inducing significant artifacts[10]. However, in practice, conversion quality is always compromised by phase incoherency in the regenerated signal, and to minimise this problem, a pitch synchronous sinusoidal model is used in our system [5], [11]. Secondly, the acoustic features which enable humans to identify speakers must be extracted and coded. These features should be independent of the message and the environment so that whatever and wherever the source speaker speaks, his/her voice characteristics can be successfully transformed to sound like the target speaker. Clearly the changes applied to these features must be capable of straightforward realization by the speech model. Thirdly, the type of conversion function and the method of training and applying the conversion function must be decided. More details on these two latter issues are presented below.

B. Spectral Parameters

As indicated above, the overall shape of the spectral envelope provides an effective representation of the vocal tract characteristics of the speaker and the formant structure of voiced sounds. Generally, there are several ways to estimate the spectral envelope, such as using LPC [12], cepstral coefficients [13] and line spectral frequencies (LSF) [15]. In Stylianou's system [4], a set of discrete MFCC coefficients is used to represent the spectral envelope. They concluded that this method provides a better envelope fit at the specified frequency points than LPC-based methods. Whilst Kain in [5] used line spectral frequencies (LSF) converted from the LPC filter parameters for the reason that LSFs have better linear interpolation attributes. Both methods have been studied in our previous research in [6] and [11]. LSF is the final choice for our system as it requires less coefficients to efficiently capture the formant structure. For cases with limited training data, this is rather crucial. Furthermore the robust interpolation properties of LSF are advantageous when using linear transformations for the conversion function.

The main steps in estimating the LSF envelope for each speech frame are as follows,

1) Use the amplitudes of the harmonics $a_k(k = 1, \cdots, K)$ determined by the pitch synchronous sinusoidal model to represent the magnitude spectrum. K is determined by the fundamental frequency F_0, its value can typically range from 50 to 200.
2) Resample the magnitude spectrum non-uniformly according to the bark scale frequency warping using cubic spline interpolation [14].
3) Compute the LPC coefficients by applying the Levinson-Durbin algorithm to the autocorrelation sequence of the warped power spectrum.
4) Convert the LPC coefficients to LSF.

In order to maintain adequate encoding of the formant structure, LSF spectral vectors with an order of $p = 15$ were used throughout our voice conversion experiments.

C. Linear Transforms

We now turn to the key problem of finding an appropriate conversion function to transform the spectral parameters. Assume that the training data contains two sets of spectral vectors $\mathbf{X}$ and $\mathbf{Y}$ which respectively encode the speech of the source speaker and the target speaker,

$$\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x}_\tau]; \quad \mathbf{Y} = [\mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_T}]; \tag{1}$$

where each vector $\mathbf{x}_i$ (or $\mathbf{y}_j$) is of dimension p.

A straightforward method to convert the source vectors is to use a linear transform. In the general case, the linear transformation of a p-dimensional vector $\mathbf{x}$ is represented by a $p \times (p+1)$ dimensional matrix W applied to the extended vector $\bar{\mathbf{x}} = [\mathbf{x}', 1]'$. Since there are a wide variety speech sounds, a single global transform is not sufficient to capture the variability in human speech. Therefore, a commonly used technique is to classify the speech sounds into classes using a statistical classifier such as a Gaussian Mixture Model (GMM) and then apply a class-specific transform. Thus, in this case, the source data set $\mathbf{X}$ would be first grouped into N classes using a GMM, and then a class-specific transform W_n would be estimated for each speech class C_n for $n = 1, \cdots, N$.

However, in practice, the selection of a single transform from a finite set of N transformations can lead to discontinuities in the output signal. In addition, the selected transform may not be appropriate for source vectors that fall in the overlap area between classes. Hence, in order to generate more robust transformations, a soft classification is preferred in which all N transformations contribute to the conversion of the source vector. The contribution degree of each transformation matrix depends on the degree to which that source vector belongs to the corresponding speech class. Thus the conversion function applied to each source vector has the following general interpolation form,

$$\mathcal{F}(\mathbf{x}) = (\sum_{n=1}^{N} \lambda_n(\mathbf{x}) W_n)\bar{\mathbf{x}} \tag{2}$$

where λ_n is the interpolation weight of transformation matrix W_n, and its value is given by the probability of vector $\mathbf{x}$ falling in speech class C_n, i.e.

$$\lambda_n(\mathbf{x}) = P(C_n|\mathbf{x}) = \frac{\alpha_n N(\mathbf{x}; \mu_n, \Sigma_n)}{\sum_{i=1}^{N} \alpha_i N(\mathbf{x}; \mu_i, \Sigma_i)} \tag{3}$$

where $\{\alpha_n\}$, $\{\mu_n\}$ and $\{\Sigma_n\}$ are the weights, means and covariances of the GMM model respectively, and $N()$ denotes the normal distribution. It should be noted that if $\lambda_n(\mathbf{x})$ is set as

$$\lambda_n(\mathbf{x}) = \begin{cases} 1 & for \;\; n = argmax\big(P(C_n|\mathbf{x})\big) \\ 0 & otherwise \end{cases} \tag{4}$$

then a hard classification is applied to the conversion function in equation (2).

The conversion function $\mathcal{F}$ is entirely defined by the $p \times (p+1)$ dimensional matrices W_n, for $n = 1, \cdots, N$. Two different estimation methods can be used to train these transformation matrices.

1) Least Square Error Estimation: When parallel training data is available, the transformation matrices can be estimated directly using the least square error (LSE) criterion. In this case, the source and target vectors are time aligned such that each source training vector $\mathbf{x}_i$ corresponds to a target training vector $\mathbf{y}_i$. For ease of manipulation, the general form of the interpolated transformation in (2) can be rewritten compactly as,

$$\begin{aligned} \mathcal{F}(\mathbf{x}) &= \Big[W_1 \vdots W_2 \vdots \cdots \vdots W_N\Big] \begin{pmatrix} \lambda_1(\mathbf{x})\bar{\mathbf{x}} \\ \cdots \\ \lambda_2(\mathbf{x})\bar{\mathbf{x}} \\ \cdots \\ \vdots \\ \cdots \\ \lambda_N(\mathbf{x})\bar{\mathbf{x}} \end{pmatrix} \\ &= \mathbf{W}\mathbf{\Lambda}(\mathbf{x}) \end{aligned} \tag{5}$$

where

$$\mathbf{W} = \Big[W_1 \vdots W_2 \vdots \cdots \vdots W_N\Big]_{p \times \big(N \times (p+1)\big)} \tag{6}$$

and

$$\mathbf{\Lambda}(\mathbf{x}) = \begin{pmatrix} \lambda_1(\mathbf{x})\bar{\mathbf{x}} \\ \cdots \\ \lambda_2(\mathbf{x})\bar{\mathbf{x}} \\ \cdots \\ \vdots \\ \cdots \\ \lambda_N(\mathbf{x})\bar{\mathbf{x}} \end{pmatrix}_{\big(N \times (p+1)\big) \times 1} \tag{7}$$

Gathering all the training vectors into single matrices $\mathbf{X}$ and $\mathbf{Y}$ as above gives the following set of simultaneous equations for estimating $\mathbf{W}$,

$$\mathbf{Y} = \mathbf{W}\mathbf{\Lambda}(\mathbf{X}) \tag{8}$$

The standard least-squares solution to equation (8) is then

$$\mathbf{W} = \mathbf{Y}\mathbf{\Lambda}(\mathbf{X})'\Big(\mathbf{\Lambda}(\mathbf{X})\mathbf{\Lambda}(\mathbf{X})'\Big)^{-1} \tag{9}$$

In practice, we use the pseudo inverse in equation (9), since for many cases where the number of mixtures is large and the amount of training data is limited, $\mathbf{\Lambda}(\mathbf{X})\mathbf{\Lambda}(\mathbf{X})'$ will become non-positive definite due to numerical errors. This LSE training approach is essentially equivalent to Stylianou's approach in [4] but with a more interpretable and flexible formulation.

The accurate alignment of source and target vectors in the training set is crucial for a robust estimation of the transformation matrices. Normally a Dynamic Time Warping (DTW) algorithm is used to obtain the required time alignment where the local cost function is the spectral distance between source and target vectors. However, the alignment obtained using this method will sometimes be distorted when the source and target speakers are very different, this is especially a problem in cross gender transformation.

Where the orthography of the training data is available, a more robust approach is to use a speech recogniser in "forced alignment mode" to find corresponding phone or sub-phone boundaries. A DTW algorithm can then be employed to align the corresponding segments between the source and target utterances. In the work described here, the HTK recogniser is used [18] with a set of speaker independent monophone HMMs. The recogniser is used to force align both the source

and the corresponding target utterance, after which the utterances can be labelled into time-marked segments where each segment corresponds to one HMM state.

2) Maximum Likelihood Estimation: As noted in the introduction, the provision of parallel training data is not always feasible and hence it would be useful if the required transformation matrices could be estimated from non-parallel data. The form of equation (5) suggests that, analogous to the use of transforms for adaptation in speech recognition [7], [8], maximum likelihood (ML) should provide a framework for doing this.

First consider the simple case of one global linear transform W and assume that there is a statistical model $\mathcal{M}$ that has been trained to well-represent the target speaker's speech. Then the optimal linear transform $\hat{W}$ applied to the source vectors $\mathbf{X} = \{\mathbf{x}_t\}$ would be the one that results in the converted vectors having maximum log likelihood with respect to the target speech model, i.e.

$$\hat{W} = \underset{W}{\text{argmax}} \sum_{t=1}^{T} logP(W\bar{\mathbf{x}}_t|\mathcal{M}) \quad (10)$$

$$= \underset{W}{\text{argmax}}\ L(W\bar{\mathbf{X}}|\mathcal{M}) \quad (11)$$

where, in our case, the statistical model $\mathcal{M}$ is a Hidden Markov Model (HMM).

There is no closed-form solution for $\hat{W}$, but an efficient iterative solution is possible using Expectation-Maximisation (EM). Consider the source data set $\mathbf{X}$ transformed at each iteration step k by $W^{(k)}$ to give a converted data set $\tilde{\mathbf{X}}^{(k)} = \{\tilde{\mathbf{x}}_t^{(k)}\}$, where $\tilde{\mathbf{x}}_t^{(k)} = W^{(k)}\bar{\mathbf{x}}_t$, (note that $k > 0$ and $\tilde{\mathbf{x}}_t^{(0)} = \mathbf{x}_t$), the log likelihood can then be decomposed as,

$$L(\tilde{\mathbf{X}}^{(k)}|\mathcal{M}) = \sum_{t=1}^{T} logP(W^{(k)}\bar{\mathbf{x}}_t|\mathcal{M})$$

$$= \sum_{t=1}^{T}\sum_{m} P(q_m(t)|\tilde{\mathbf{x}}_t^{(k-1)},\mathcal{M})\ logP(W^{(k)}\bar{\mathbf{x}}_t|\mathcal{M}) \quad (12)$$

$$= \sum_{t=1}^{T}\sum_{m} P(q_m(t)|\tilde{\mathbf{x}}_t^{(k-1)},\mathcal{M})\ log\frac{P(\tilde{\mathbf{x}}_t^{(k)}, q_m(t)|\mathcal{M})}{P(q_m(t)|\tilde{\mathbf{x}}_t^{(k)},\mathcal{M})}$$

$$= \mathcal{Q}(\tilde{\mathbf{X}}^{(k-1)},\tilde{\mathbf{X}}^{(k)}) - \mathcal{K}(\tilde{\mathbf{X}}^{(k-1)},\tilde{\mathbf{X}}^{(k)}) \quad (13)$$

where

$$\mathcal{Q}(\tilde{\mathbf{X}}^{(k-1)},\tilde{\mathbf{X}}^{(k)}) = \sum_{t=1}^{T}\sum_{m} P(q_m(t)|\tilde{\mathbf{x}}_t^{(k-1)},\mathcal{M})\ logP(\tilde{\mathbf{x}}_t^{(k)}, q_m(t)|\mathcal{M}) \quad (14)$$

$$\mathcal{K}(\tilde{\mathbf{X}}^{(k-1)},\tilde{\mathbf{X}}^{(k)}) = \sum_{t=1}^{T}\sum_{m} P(q_m(t)|\tilde{\mathbf{x}}_t^{(k-1)},\mathcal{M})\ logP(q_m(t)|\tilde{\mathbf{x}}_t^{(k)},\mathcal{M}). \quad (15)$$

Here $q_m(t)$ indicates Gaussian component m of the target HMM $\mathcal{M}$ at time t, and the sum is taken over all components which can be aligned with $\mathbf{x}_t$. Hence $\sum_m P(q_m(t)|\tilde{\mathbf{x}}_t^{(k-1)},\mathcal{M}) = 1$ which justifies the expansion in equation (12).

Noting that the likelihood in equation (13) only depends on the second parameter of $\mathcal{Q}$ and $\mathcal{K}$, it follows that

$$L(\tilde{\mathbf{X}}^{(k-1)}|\mathcal{M}) = \mathcal{Q}(\tilde{\mathbf{X}}^{(k-1)},\tilde{\mathbf{X}}^{(k-1)}) - \mathcal{K}(\tilde{\mathbf{X}}^{(k-1)},\tilde{\mathbf{X}}^{(k-1)}) \quad (16)$$

and by Jensen's Inequality,

$$\mathcal{K}(\tilde{\mathbf{X}}^{(k-1)},\tilde{\mathbf{X}}^{(k)}) \leq \mathcal{K}(\tilde{\mathbf{X}}^{(k-1)},\tilde{\mathbf{X}}^{(k-1)}). \quad (17)$$

Hence if the auxiliary function $\mathcal{Q}(\tilde{\mathbf{X}}^{(k-1)},\tilde{\mathbf{X}}^{(k)})$ is maximised such that $\mathcal{Q}(\tilde{\mathbf{X}}^{(k-1)},\tilde{\mathbf{X}}^{(k)}) \geq \mathcal{Q}(\tilde{\mathbf{X}}^{(k-1)},\tilde{\mathbf{X}}^{(k-1)})$, then it follows from equations (13), (16) and (17) that $L(\tilde{\mathbf{X}}^{(k)}|\mathcal{M}) \geq L(\tilde{\mathbf{X}}^{(k-1)}|\mathcal{M})$. Thus, repeated maximisation of equation (14) to find $W^{(k)}$, each time updating the Gaussian component occupation probabilities to use the previous transform, leads eventually to $\hat{W}$. In practice, it is found that convergence occurs quickly and only a few iterations are required. Indeed, often just one iteration is sufficient for similar speakers.

The required maximisation at each step k proceeds by rewriting the auxiliary function in (14) (with the constant terms suppressed) as,

$$\mathcal{Q}(\tilde{\mathbf{X}}^{(k-1)},\tilde{\mathbf{X}}^{(k)}) = -\frac{1}{2}\sum_{t=1}^{T}\sum_{m}\beta_m(t) \Big[(W\bar{\mathbf{x}}_t - \mu_m)'\Sigma_m^{-1}(W\bar{\mathbf{x}}_t - \mu_m)\Big] \quad (18)$$

where $W = W^{(k)}$ is the transform at step k, and μ_m and Σ_m are the mean vector and covariance matrix of Gaussian component m in $\mathcal{M}$, and $\beta_m(t)$ is,

$$\beta_m(t) = \beta_m^{(k-1)}(t) = P(q_m(t)|\tilde{\mathbf{x}}_t^{(k-1)},\mathcal{M}). \quad (19)$$

Note that the initial value of $\beta_m^{(0)}(t) = P(q_m(t)|\mathbf{x}_t,\mathcal{M})$.

Differentiating $\mathcal{Q}$ in equation (18) with respect to W and equating to zero gives,

$$\sum_{t=1}^{T}\sum_{m}\beta_m(t)\Sigma_m^{-1}\mu_m\bar{\mathbf{x}}_t' = \sum_{t=1}^{T}\sum_{m}\beta_m(t)\Sigma_m^{-1}W\bar{\mathbf{x}}_t\bar{\mathbf{x}}_t'. \quad (20)$$

The left-hand side of equation (20) is independent of W so call this $\mathbf{Z}$. Introducing variables,

$$\mathbf{V}^{(t)} = \sum_{m}\beta_m(t)\Sigma_m^{-1} \quad (21)$$

$$\mathbf{D}^{(t)} = \bar{\mathbf{x}}_t\bar{\mathbf{x}}_t' \quad (22)$$

equation (20) can then be rewritten as

$$\mathbf{Z} = \sum_{t=1}^{T}\mathbf{V}^{(t)}W\mathbf{D}^{(t)}. \quad (23)$$

Assuming that $\mathcal{M}$ has diagonal covariance matrices, a closed form solution can be derived by defining a new matrix $\mathbf{G}^{(i)}$ with elements[7],

$$g_{jq}^{(i)} = \sum_{t=1}^{T} v_{ii}^{(t)}d_{jq}^{(t)} \quad j,q = 1,\cdots,(d+1) \quad (24)$$

W is then calculated row by row using

$$\mathbf{w}_i' = \mathbf{G}^{(i)-1}\mathbf{z}_i' \tag{25}$$

where $\mathbf{w}_i$ and $\mathbf{z}_i$ are the i-th row of W and $\mathbf{Z}$, respectively.

To estimate multiple transforms using this scheme, a source GMM is used to assign the source vectors to classes via equation (4) as in the LSE estimation scheme. A transform matrix is then estimated separately for each class using the above ML scheme applied to just the data for that class. Though it is theoretically possible to estimate multiple transforms using soft classification, in practice, matrices $\mathbf{D}$ and $\mathbf{G}$ will become too large to invert. Hence the simpler hard classification approach is used here.

As with the least mean squares method using parallel data, performance is greatly improved if sub-phone segment boundaries can be accurately determined in the source data using the target HMM $\mathcal{M}$ and "forced alignment" recognition mode. This enables the set of Gaussians evaluated for each source frame to be limited to just those associated with the HMM state corresponding to the associated sub-phone. This does, of course, require that the orthography of the source utterances be known. Similarly, knowing the orthography of the target training data makes training the target HMM simpler and more effective. More details on implementation issues are given in the following subsection.

D. Evaluation

1) Data: The VOICES database from OGI is used for evaluation[5]. This corpus contains recorded speech from 12 different speakers reading 50 phonetically rich sentences. Each sentence is spoken 3 times by each speaker. The speech data was recorded at 22K Hz sampling rate using a 16 bit encoding in a professional sound-booth with high quality headphones. The recording procedure involved a "mimicking" approach which resulted in a high degree of natural 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 speaker-pair, the first 120 utterances are used as training data, and the remaining 30 utterances form the test set.

2) Objective Measure: Objective measures seek to evaluate the differences between two speech signals. Since many perceived sound differences can be interpreted in terms of differences of spectral features [16], spectral distortion is considered to be a reasonable metric both mathematically and subjectively. In speech processing, a log spectral measure is often used to determine the distance between two spectra [17]. Similarly in this paper, the log spectral distortion between two spectral envelopes was used to provide an objective measure of the conversion performance

$$d(S_1, S_2) = \frac{1}{K}\sum_{k=1}^{K}(10log_{10}a_k^1 - 10log_{10}a_k^2)^2 \tag{26}$$

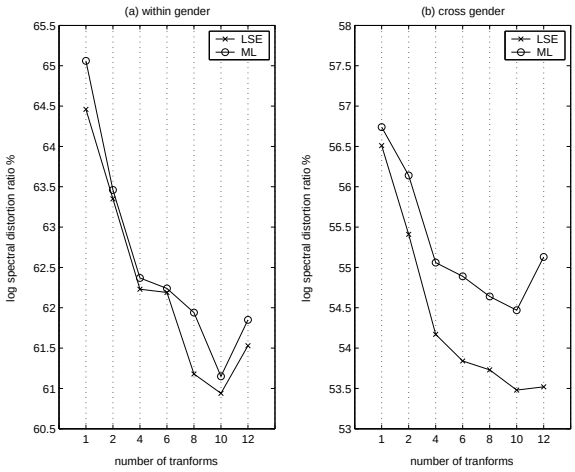

Fig. 1. *Spectral distortion ratio for LSE and ML transforms, (a) within-gender voice conversion, (b) cross-gender voice conversion.*

where $\{a_k\}$ are the amplitudes resampled from the normalised spectral envelope S at K uniformly spread frequencies, and K is set to 100 throughout our experiments. A distortion ratio is then used to compare the converted-to-target distortion with the source-to-target distortion, which is defined as,

$$D = \frac{\sum_{t=1}^{L} d(S_{tgt}(t), S_{conv}(t))}{\sum_{t=1}^{L} d(S_{tgt}(t), S_{src}(t))} \times 100\% \tag{27}$$

where $S_{tgt}(t)$, $S_{src}(t)$ and $S_{conv}(t)$ are the target spectral envelope, source spectral envelope and the converted spectral envelope at time t respectively. The summation in each case is computed over time-aligned data and L is the total number of test vectors after time alignment.

It should be noted that since the spectral distortion also depends on the degree to which the time-alignment process can align similar vectors, it is typically quite large, even when applied to the same speaker. For example, the average log spectral distortion d between two utterances with identical content and spoken by the same speaker can vary from 5 to 10 dB, whilst the distortion between two different speakers would normally be in the range from 13 dB to 20 dB. So in practice, a distortion ratio of $D = 50\%$ would represent acceptable conversion performance. Note also that a 100% distortion ratio corresponds to the distortion between the source and target spectrum.

3) LSE and ML Comparison: The training of LSE transforms is straightforward. First, a GMM model is trained on the source vectors and the interpolation weights are computed according to equation (3). Second, a forced alignment of all utterances is computed and sub-phone boundaries are marked. Third, DTW-based time alignment is applied constrained by these sub-phone boundaries to produce a set of aligned source-target vector pairs. In the case of the OGI Voice corpus, around 30,000 vector pairs are obtained for each speaker pair. Once the training data has been extracted, the transformation matrices can be computed using equation (9).

The ML training scheme is a little more complex. First, the orthography of the target speaker's training data is known

and used to train a monophone HMM set with 4 Gaussian mixture components per state. Since the data is sparse, a tied-mixture technique is employed such that the HMM states share Gaussian components but with different weights for different states. The same source GMM as for LSE is used to classify the source vectors so that multiple ML transforms can be estimated. As suggested above, the source utterances were force-aligned to map every source training vector to a specific HMM sub-phone state, which therefore required that the orthography of the source training data is also known. The Gaussian component occupation probabilities were then computed as per equation (19) and then the required transformation matrices estimated using equation (25).

The number of iterations required depends on the source and target data. One iteration is typically sufficient for within-gender conversion. However, for cross-gender conversion, two or more iterations are necessary.

Fig.1 shows the spectral distortion ratio using LSE and ML transforms. For both methods, the distortion decreases as the number of transforms is increased until data sparsity results in over-training. For these experiments with approximately 30,000 training vectors per speaker, the results suggest that around 10 transforms is optimal. This corresponds to $10 \times (15 \times 16) = 2400$ parameters. The difference between LSE and ML transforms in the within gender voice conversion is very small as shown by Fig.1(a), however the difference is larger for the cross gender conversion case as shown in Fig. 1(b). However, defining the signal-to-noise ratio between the LSE and ML transformed utterances as

$$SNR = 10 \times log_{10} \frac{\sum_{n=1}^{N} \mathbf{s}_{lse}(n)^2}{\sum_{n=1}^{N} [\mathbf{s}_{lse}(n) - \mathbf{s}_{ml}(n)]^2} \qquad (28)$$

Table I shows that the signal to noise ratio is actually very high even in the cross-gender case and should be imperceptable to human listeners. To test this further, a formal listening test was conducted whereby listeners were presented with pairs of utterances generated by the LSE and the ML method respectively, and asked to select the one with the highest perceived quality. Note that in this experiment only the quality of the converted speech is of concern, not the transformation accuracy of speaker identity. The latter aspect is evaluated in section IV. Table II indicates that the listeners show almost equal preference for the ML and LSE converted utterances and a two tailed t-test indicates that the difference is indeed insignificant (p=0.499 in support of the null hypothesis).

TABLE I

The SNR ratio in dB between LSE and ML transformed utterances.

	within gender	cross gender
SNR	30.4	24.1

Note that although the distortion ratio of the cross gender conversion seems much lower than that of the within gender conversion as shown in Fig. 1, the average log spectral distortion value is actually higher (8.83 dB for cross gender and 8.37 dB for within gender). This is simply because the source to target distortion of cross gender conversion is much larger than that of within gender conversion.

TABLE II

The result of a preference test to compare LSE and ML transformed utterances.

	ML	LSE
preference	48.3%	51.7%

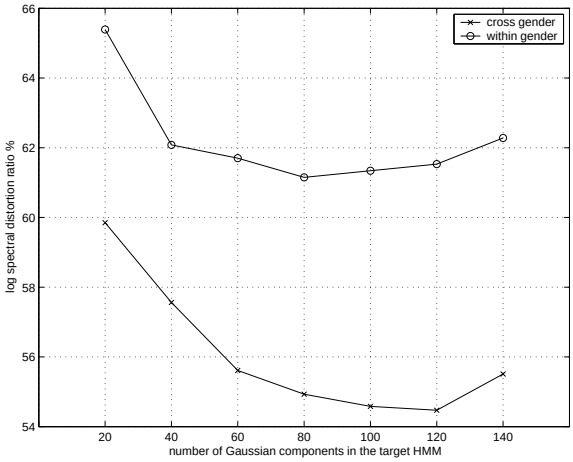

Fig. 2. *Spectral distortion ratio over different numbers of Gaussian components in the target HMM. (10 ML transforms)*

Although the differences are small and are subjectively imperceptable, distortion is nevertheless consistently lower in all cases for LSE derived transforms compared to ML derived transforms. This may be because the use of time-aligned parallel data in the LSE case allows the evolution of the spectral vectors to be captured whereas in the ML case, the spectral evolution is only approximately modelled by the HMM state transitions. This suggests that improving the modelling accuracy of the target HMM should improve the ML transforms.

Fig. 2 shows that increasing the number of Gaussian components in the target HMM can reduce the spectral distortion ratio, however this is limited by data sparsity. Fig. 3 shows that when increasing the number of training vectors, the spectral distortion ratio decreases for both the ML and LSE cases. Thus, not surprisingly, both methods can benefit from more training data but the ML method can benefit from having more target training data even when the source data is limited. This latter point can be important for applications where there is a very large amount of data available for the target but only limited data for the source [21].

TABLE III

The spectral distortion ratios of LSE, parallel ML and non-parallel ML transforms.

	LSE	parallel ML	non-parallel ML
within gender	65.1%	67.4%	68.0%
cross gender	57.1%	61.8%	61.1%

The above evaluation was conducted using entirely parallel training data in order to be able to compare the LSE and

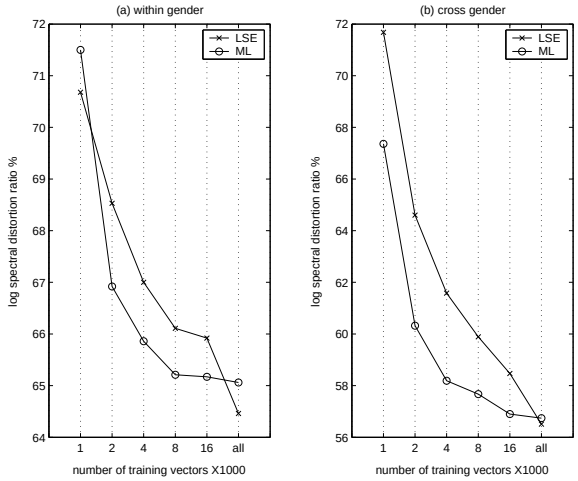

Fig. 3. *Spectral distortion ratio for single LSE and ML transforms over different numbers of training vectors, (a) within gender voice conversion, (b) cross gender voice conversion.*

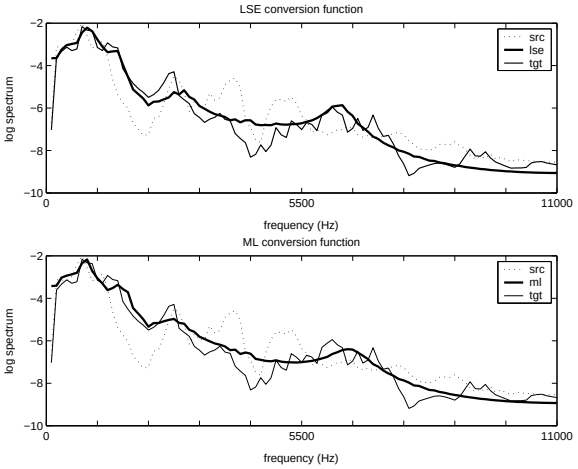

Fig. 4. *Examples of spectral envelope conversion using ML and LSE estimated linear transforms. (a) Spectral envelope conversion using LSE estimated transforms. (b) Spectral envelope conversion using ML estimated transforms. (dotted line: the source spectral envelope; lighter solid line: the target spectrum; dark solid line: the converted spectral envelope.)*

ML approaches. However, the use of parallel data for the ML approach may flatter the results compared to what would have been obtained with truly non-parallel training data. To test this a further experiment was conducted in which the 120 training utterances for each speaker were divided into two equal sets. For the LSE estimation, the first 60 utterances of both source and target speaker were used for training. For ML estimation, however, the first 60 utterances of the target speaker were used to train the tied-mixture HMM. Then both sets of source utterances were used to generate two different ML transforms: the "parallel" ML transforms, and the "non-parallel" ML transforms. Since the training data was only half the size of the previous experiments, only 4 transforms were estimated in each case. As shown in Table III, the parallel ML and the non-parallel ML transforms gave very similar performance, although both are worse than the LSE transforms. The latter is almost certainly because the target HMM was badly undertrained with only 60 utterances.

Finally, an example of spectral envelope conversion using LSE and ML transforms is shown in Fig. 4. Both methods have converted the source spectral envelope to match the target, however many spectral details have been lost and this is a major cause of the spectral distortion. Moreover, listeners report that overall the converted speech is not high quality with many artifacts including a muffled effect. In the following section, these artifacts are analysed and solutions presented.

III. System Enhancement

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

A. Phase Prediction

As is well known, the spectral magnitude and phase of human speech are highly correlated. In the baseline system, when only spectral magnitudes are modified and the original phase is preserved, a harsh quality is introduced into the converted speech. However, to simultaneously model the magnitude and phase 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 $(C_1, \cdots, C_M)$ such as in the ML estimation. Then for each target envelope $\mathbf{y}_t$ we have a set of posterior probabilities $P(C_m|\mathbf{y}_t)$. The vector $\mathcal{P}(\mathbf{y}_t)$ composed from these probabilities can then be regarded as another form of representation of the spectral shape,

$$\mathcal{P}(\mathbf{y}_t) = [P(C_1|\mathbf{y}_t), \cdots, P(C_M|\mathbf{y}_t)]' \tag{29}$$

Each element $P(C_i|\mathbf{y}_t)$ of this vector can be regarded as the weight of a codebook entry S_i and the set of M codebook entries

$$\mathcal{T} = [S_1, \cdots, S_M] \tag{30}$$

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,

$$E = \sum_{t=1}^{N} (s(t) - \mathcal{T}\mathcal{P}(\mathbf{y}_t))'(s(t) - \mathcal{T}\mathcal{P}(\mathbf{y}_t)) \tag{31}$$

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 (31) is then

$$\mathcal{T} = \Big(\sum_{t=1}^{N} s(t)\mathcal{P}(\mathbf{y}_t)'\Big)\Big(\sum_{t=1}^{N} \mathcal{P}(\mathbf{y}_t)\mathcal{P}(\mathbf{y}_t)'\Big)^{-1} \tag{32}$$

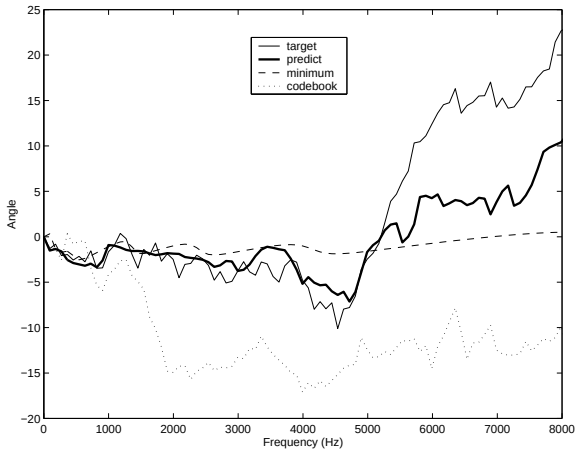

Fig. 5. *Example of the unwrapped phase spectra generated by minimum phase, phase codebook and phase prediction. Light solid line: the target phase spectrum. Dark solid line:the phase spectrum generated by phase prediction. Dashed line: the phase spectrum generated by minimum phase. Dotted line: the phase spectrum generated by phase codebook.*

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

$$\tilde{s}(t) = \mathcal{T}\mathcal{P}(\tilde{\mathbf{x}}_t); \tag{33}$$

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.

This phase prediction method has been compared with two other popular phase coding methods: the minimum phase and the phase codebook approach[19]. The experiments were conducted as follows. First the original signal was analyzed using the pitch synchronous sinusoidal model, then the original phase spectra were replaced by the synthetic phase spectra generated respectively by the minimum phase method, the phase codebook method and the phase prediction method. In our experiments, the number of speech classes M for each speaker is 64, depending on the number of training vectors that can be obtained. Additionally, to reduce the modelling error, the pitch synchronous sinusoidal model used in our experiments has automatically adjusted the end points of each pitch period to be positioned at the zero-crossing points, and each speech frame was normalized by energy before modelling the phases. Fig. 5 shows an example of the unwrapped phase spectra generated by the minimum phase method, phase codebook and the phase prediction method. Clearly, the phase prediction spectra more closely fits the target phase spectra. Table IV shows the signal to noise ratio (SNR) using the above three different phase coding methods. The phase prediction approach outperforms the other two approaches and furthermore the improvement in audio quality is noticeable in listening tests.

TABLE IV

The SNR ratio in dB of three different phase coding methods.

minimum phase	codebook phase	phase prediction
5.7	12.3	14.4

B. Spectral Refinement

As noted earlier in Fig.4, although the formant structure of the source speech has been transformed to match the target, the spectral detail has been lost as a result of reducing the dimensionality of the envelope representation during the transform. Another clearly visible effect is the broadening of the spectral peaks caused, at least in part, by the averaging effect of the estimation method. All these degradations lead to muffled effects in the converted speech.

To solve this problem, a straightforward idea is to re-introduce the lost spectral details to the converted envelopes. A spectral residual prediction approach has been developed to do this based on the residual codebook method proposed in [5], where the codebook is trained using a GMM model.

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

$$r_t = 20log_{10}H(t)_{sin} - 20log_{10}H(t)_{env} \tag{34}$$

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. In our experiments, r_t is a 100 dimensional vector resampled from the residual contour. Each spectral residual r_t is associated with an LSF vector $\mathbf{y}_t$, and is therefore associated with a set of posterior probabilities as in equation (29). Similar to the phase prediction approach, a residual codebook $\mathcal{R} = [R_1, R_2, \cdots, R_M]$ is trained. The prediction error on the training data is defined as follows,

$$\mathcal{E} = \sum_{t=1}^{T}(r_t - \mathcal{R}\mathcal{P}(\mathbf{y}_t))'(r_t - \mathcal{R}\mathcal{P}(\mathbf{y}_t)) \tag{35}$$

and the solution to $\mathcal{R}$ is

$$\mathcal{R} = \Big(\sum_{t=1}^{T} r_t\mathcal{P}(\mathbf{y}_t)'\Big)\Big(\sum_{t=1}^{T}\mathcal{P}(\mathbf{y}_t)\mathcal{P}(\mathbf{y}_t)'\Big)^{-1} \tag{36}$$

After the residual codebook $\mathcal{R}$ is obtained, the spectral residual needed to compensate each converted spectral envelope can be predicted straightforwardly based on the posterior probabilities.

TABLE V

Effect of residual prediction as measured by log spectral distortion ratios computed over the real spectrum.

	within gender	cross gender
before RP	74.4%	73.0%
after RP	54.3%	53.8%

Table V shows the log spectral distortion ratio before and after residual prediction (RP). Here the log spectral distortion was computed over the real spectrum instead of the spectral envelope. As can be seen, the use of residual prediction results in a 20% absolute decrease in the spectral distortion ratio for both cross and within gender conversions.

As mentioned earlier, transform-based voice conversion systems have a tendency to broaden the formants in the converted speech. To mitigate this effect and suppress noise in the spectral valleys, a further spectral refinement is to apply

a perceptual filter to the regenerated spectral envelope of all voiced sounds. The perceptual filter is defined as,

$$H(\omega) = \frac{A(z/\beta)}{A(z/\gamma)}, 0 < \gamma < \beta \leq 1 \quad (37)$$

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 [20] and its more general use in voice conversion is discussed in [6].

C. 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 from the source. 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 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, the target training data is first labelled using the forced alignment technique mentioned in the ML estimation scheme, so that each speech frame is given an HMM state label together with a voiced/unvoiced decision. All these labels and the target speech frames are then gathered together into a database.

When a segment of unvoiced speech from the source speaker needs to be transformed, each frame in the segment is first labelled with its corresponding HMM state using the same forced alignment technique. 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 HMM state 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.

IV. Evaluation of Enhanced System

In order to test the overall subjective quality of the voice morphing system, 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 VI gives the percentage of the converted utterances that were labelled as closer to the target for each case, where the "baseline system" refers to the system that only transforms the spectral envelopes and "enhanced system" refers to the system that integrates all of the refinements described in section III. The results clearly show that the enhanced 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.

TABLE VI
Results from the ABX test.

	baseline system	enhanced system
ABX	86.4%	91.8%

To assess speech quality between the baseline system and the enhanced system, a second preference test was conducted whereby listeners were presented with pairs 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 VII indicates that most listeners prefer the converted speech generated by the enhanced system. Moreover, as the p-value of this t-test is 0.023, much lower than the significance level 0.05, the difference between the enhanced system and the baseline system in Table VII is statistically significant. This is consistent with the previous objective evaluations. Although the relative contribution of each individual refinement is very difficult to measure, informal tests suggest that the spectral refinement described in Section III B above contributes the most to quality enhancement.

TABLE VII
Results from the preference test.

	baseline system	enhanced system
preference	38.9%	61.1%

V. Conclusion

This paper has presented a study of voice morphing based on interpolated linear transformations. The study has focussed on two main issues. Firstly, a Maximum Likelihood method of estimating the required transformation functions has been developed which does not depend on the availability of parallel training data. Comparative tests have shown that this method

is equal in performance to least mean square estimators using parallel data however it is much more flexible. Secondly, the main causes of artifacts in the converted speech have been identified as excessive spectral smoothing, unnatural phase prediction and conversion of unvoiced speech. Solutions to these problems have been proposed and shown to be effective using a variety of objective and subjective measures.

Overall, the results show that transform-based voice conversion can produce the required identity change whilst maintaining acceptable quality. In particular, the flexibility of the ML training technique combined with the described quality enhancements offer the promise of immediate application in telephone-based applications such as customising voice output, novelty voice-messaging, etc.

Nevertheless, there is still considerable scope for further work. The most serious weakness in the current system is the prosodic modelling. Shifting and scaling the pitch to match the mean and variance of the target speaker is only adequate when the speakers are similar. When the speakers are very different (e.g. when converting a British English speaker to an American English speaker), the resulting perception of identity is ambiguous. Also, although the enhancements described in this paper give a substantial improvement in overall audio quality, there is still residual distortion making it unsuitable for applications where "studio quality" is required in the converted speech.

ACKNOWLEDGMENT

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

REFERENCES

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

[2] L. Arslan, D. Talkin, *"Speaker Transformation Algorithm using Segmental Codebooks (STASC)"*, Speech Communication, 1999.

[3] C.-H. Ho, D. Rentzos, S. Vaseghi, *"Formant Model estimation and transformation for Voice Morphing"*, Proc. ICSLP, 2002.

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

[5] A. Kain, *"High resolution voice transformation"*, PhD dissertation, OGI, 2001.

[6] H. Ye and S. Young, *"Perceptually Weighted Linear Transformation for Voice Conversion"*, Eurospeech 2003.

[7] C.J. Leggetter and P.C. Woodland, *"Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov model"*. Computer Speech and Language, vol. 9, pp. 171-185, 1995.

[8] M.J.F. Gales, *"Maximum Likelihood Linear Transformations for HMM-based Speech Recognition"*. Computer Speech and Language, vol. 12, 1998.

[9] E. B. George and M. J. T. Smith, *"Speech Analysis/synthesis and Modification Using an Analysis-by-synthesis/overlap-add Sinusoidal Model"*, IEEE Trans. on Speech and Audio Proc., vol. 5, no. 5, pp. 389-406, September 1997.

[10] T. F. Quatieri and R. J. McAulay, *"Shape Invariant Time-scale and Pitch Modification of Speech"*, IEEE Trans. on Signal Proc., vol. 40, pp. 497-510, March 1992.

[11] H. Ye and S. Young, *"High Quality Voice Morphing"*, In Proceedings of ICASSP 2004.

[12] J. Wouter and W. Macon, *"Control of Spectral Dynamics in Concatenative Speech Synthesis"*, IEEE Trans. on Speech and Audio Proc., vol. 9, no. 1, pp. 30-38, January 2001.

[13] O. Cappe, J. Laroche and E. Moulines, *"Regularized estimation of cepstrum envelope from discrete frequency points"*, In Proceedings of IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, New York, 1995.

[14] M. Unser, A. Aldroubi and M. Eden, *"B-Spline Signal Processing"*, IEEE Trans. on Signal Proc., vol. 41, no. 2, pp. 821-848, February 1993.

[15] F. Itakura, *"Line Spectrum Representation of Linear Predictive Coefficients."*, J Acoust Soc Am, vol. 57, no. 4, pp. 535, 1975

[16] Rabiner, L.R. and B. H. Juang, *"Fundamental of Speech Recognition"*, Prentice hall, 1993

[17] A. Gray and J.D. Markel, *"Distance measures for speech processing"*, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-24, no.5, pp.380-391, October 1976.

[18] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland. *"The HTK Book V3"*, Cambridge University, 2000

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

[20] J.H. Chen and A. Gersho, *"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.

[21] H. Ye and S. Young, *"Voice Conversion for Unknown Speakers"*, in Proc ICSLP, Jeju, Korea, 2004.