# Perceptually Weighted Linear Transformations for Voice Conversion

*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 conversion is a technique for modifying a source speaker's speech to sound as if it was spoken by a target speaker. A popular approach to voice conversion is to apply a linear transformation to the spectral envelope. However, conventional parameter estimation based on least square error optimization does not necessarily lead to the best perceptual result. In this paper, a perceptually weighted linear transformation is presented which is based on the minimization of the perceptual spectral distance between the voices of the source and target speakers. The paper describes the new conversion algorithm and presents a preliminary evaluation of the performance of the method based on objective and subjective tests.

## 1. Introduction

Voice conversion, whose purpose is to transform a source speaker's speech to sound as if it was produced by a target speaker, 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 any voice conversion system. Various approaches have been proposed for effecting the transformation including codebook mapping [1] and linear transformations. Of these, the linear transformation technique has been shown by Stylianou et al. [2] and Kain [3] to outperform other approaches in terms of speech quality.

The spectral envelope itself can be parameterized in a number of ways, including cepstral coefficients, line spectral frequencies (LSF), b-splines, etc. However, minimizing a least square error criteria is not necessarily optimal since, as is well-known, the human auditory system is insensitive to errors near a strong tone. Therefore, it is reasonable to argue that a better way of training a transformation would be to minimize the perceptual error instead of the squared error.

In this paper, we present a perceptually weighted linear transformation approach for voice conversion. This technique applies a perceptual weighting filter to the spectral error, from which an optimized linear transformation can be trained. In section 2, a baseline voice conversion system based on linear transformation is described, and then in section 3, its performance is reported. Section 4 then describes the new transformation scheme utilising perceptual weights. The performance of the new scheme is then presented in section 5 before presenting overall conclusions in section 6.

## 2. Voice Conversion Baseline Framework

Our voice conversion system uses a pitch synchronous harmonic model for speech signal representation and modification. This model supports modifications to both the prosody and the spectral characteristics of the source signal without inducing significant artifacts[4]. To transform the spectral envelope, we use an interpolated linear transform. This is essentially equivalent to the Continuous Probabilistic Transformation proposed by Stylianou et al. [2]. Here we present it in a maximum likelihood framework to simplify the later incorporation of a perceptual weighting.

Assume that the training set contains two sets of time-aligned parallel spectral vectors $\mathbf{X}$ and $\mathbf{Y}$ where each vector $\mathbf{x}_i$ (or $\mathbf{y}_j$) is of dimension $q$.

$$\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_T}]\ \mathbf{Y} = [\mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_T}] \qquad (1)$$

The data set $\mathbf{X}$ comes from the source and $\mathbf{Y}$ comes from the target speaker. Both speak the same utterances and the time alignment is achieved by Dynamic Time Warping.

The source spectral vectors, $\mathbf{X}$, are first grouped into $M$ classes using a Gaussian mixture model (GMM). A straightforward method of speech conversion would then be to train a linear transformation for each class. Thus in this case there would be $M$ different linear transformation matrices, such that matrix $W_m$ $(m = 1, \cdots, M)$ would be used to transform all vectors in the $m$-th speech class. However, the selection of a single transform from a finite set of $M$ transformations will 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.

A more robust transformation is obtained if all $M$ transformations contribute to the conversion of each source vector. The weight of each transformation matrix then depends on the probability that the source vector belongs to the corresponding speech class. Thus the conversion function that will apply to the source vector $\mathbf{x}$ is defined by the following interpolation,

$$\mathcal{F}(\mathbf{x}) = (\sum_{m=1}^{M} \lambda_m(\mathbf{x}) W_m)\bar{\mathbf{x}} \qquad (2)$$

where $\bar{\mathbf{x}} = [\mathbf{x}', 1]'$ is the extended vector of $\mathbf{x}$ and $\lambda_m$ is the interpolation weight of matrix $W_m$, its value is given by the probability of vector $\mathbf{x}$ falling in the $m$-th speech class, i.e. $\lambda_m(\mathbf{x}) = P(C_m|\mathbf{x})$ as in [2].

The conversion function $\mathcal{F}$ is entirely defined by the $q \times (q+1)$ dimensional matrices $W_m$, for $m = 1, \cdots, M$. Equation (2) can be rewritten compactly as,

$$\mathcal{F}(\mathbf{x}) = \left[W_1 \vdots W_2 \vdots \cdots \vdots W_M\right] \begin{pmatrix} \lambda_1(\bar{\mathbf{x}})\bar{\mathbf{x}} \\ \cdots \\ \lambda_2(\bar{\mathbf{x}})\bar{\mathbf{x}} \\ \cdots \\ \vdots \\ \cdots \\ \lambda_M(\bar{\mathbf{x}})\bar{\mathbf{x}} \end{pmatrix} = \bar{\mathbf{W}}\Lambda(\mathbf{x}) \quad (3)$$

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

$$\mathbf{Y} = \bar{\mathbf{W}}\Lambda(\mathbf{X}) \quad (4)$$

The standard least-squares solution to equation (4) is then

$$\bar{\mathbf{W}} = \mathbf{Y}\Lambda(\mathbf{X})'\left(\Lambda(\mathbf{X})\Lambda(\mathbf{X})'\right)^{-1} \quad (5)$$

In our baseline system, the spectral vectors are a set of cepstral coefficients representing the spectral envelope which passes through the harmonic amplitudes. The steps used to estimate these coefficients are as follows,

1. normalize the harmonic amplitudes according to the frame energy then convert the harmonic amplitudes to the log domain.
2. use cubic spline interpolation [5] to resample the log spectral envelope at 120 frequencies. These frequencies are uniformly spaced in the mel frequency domain. The resulting set of amplitudes are denoted $a_k(k = 1, \cdots, 120)$.
3. compute the cepstral coefficients $c_i(i = 1, \cdots, d)$ by minimizing the following least squares criterion [6]

$$E = \sum_{k=1}^{120}(loga_k - log|S(\omega_k)|)^2, \omega_k = \frac{k\pi}{120} \quad (6)$$

where

$$log|S(\omega)| = c_0 + 2\sum_{i=1}^{q} c_i cos(i\omega) \quad (7)$$

In practice it is not necessary to transform the whole frequency space since relatively little of the speaker identity is carried by the speech signal at frequencies above 6kHz. Therefore, in our system the cepstral coefficients are designed to capture the spectral envelope from 0 to 6kHz. Moreover it should be noted that only the voiced speech is transformed.

## 3. Baseline Evaluation

### 3.1. Speech Corpus

The VOICES database from OGI [3] is used for evaluation. 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 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 the training data, and the remaining 30 utterances as the test set. As noted earlier, a DTW algorithm was used to align the corresponding utterances before training and testing.

### 3.2. Objective Performance Measure

A log-spectral distortion measure was used to provide an objective measure of baseline performance. This is defined as

$$d(S_1, S_2) = \sum_{k=1}^{120}(loga_k^1 - loga_k^2)^2 \quad (8)$$

where $\{a_k\}$ are the amplitudes resampled from the spectral envelope $S$ as described in section 2. Thus the overall transformation performance can be evaluated by comparing the converted-to-target distortion with the source-to-target distortion, which was defined as,

$$D = 10log_{10}\frac{\sum_{t=1}^{N} d(S_{tgt}(t), S_{cov}(t))}{\sum_{t=1}^{N} d(S_{tgt}(t), S_{src}(t))} \quad (9)$$

where $S_{tgt}(t)$,$S_{src}(t)$ and $S_{cov}(t)$ are the target spectral envelope, source spectral envelope and the converted spectral envelope at time $t$ respectively. $N$ is the total number of test vectors.

### 3.3. Experiments

The overall transformation performance was evaluated based on the log-spectral distortion over different dimensions of cepstral vectors and different number of GMM components. As shown in Fig.1, the log-spectral distortion decreased constantly as the number of GMM components is increased until the transformation function was overtrained such as in the 25 dimension case with 32 GMM components. Also, it appears that increasing the vector dimension also helps to improve the transformation performance. However both of these improvements are constrained by the amount of available training data (around 30,000 pairs of vectors per task).

After changing the spectral envelope of the source speech, the pitch values of the voiced frames were scaled to match the target prosody, and then the final converted speech was provided to a set of listeners to evaluate the subjective performance. These informal listening tests suggested that although the speaker identity was successfully transformed, the final audio quality had a muffled effect. This probably arises because the least square error estimation offers equal weights to the spectral differences at different frequencies and this will result in a flattened converted spectral envelope. Thus it is reasonable to anticipate that a perceptually weighted distance which magnifies the spectral difference at the formants or spectral peaks and minimizes the difference at spectral valleys will be a better alternative.

## 4. Perceptually Weighted Linear Transformation

The perceptually weighted log-spectral distance between two spectral envelopes is defined as

$$\tilde{d} = (A_1 - A_2)'P(A_1 - A_2) \quad (10)$$

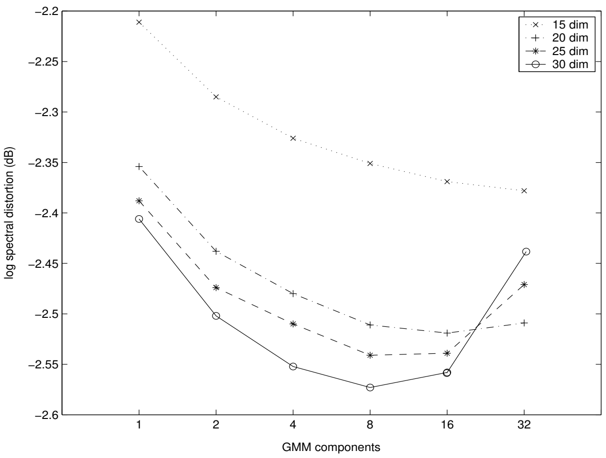


Figure 1: *Log spectral distortion over different GMM components and different dimensions of cepstral vectors (15, 20, 25 dim). The 0dB value corresponds to the initial distortion between the source and the target spectral envelope.*

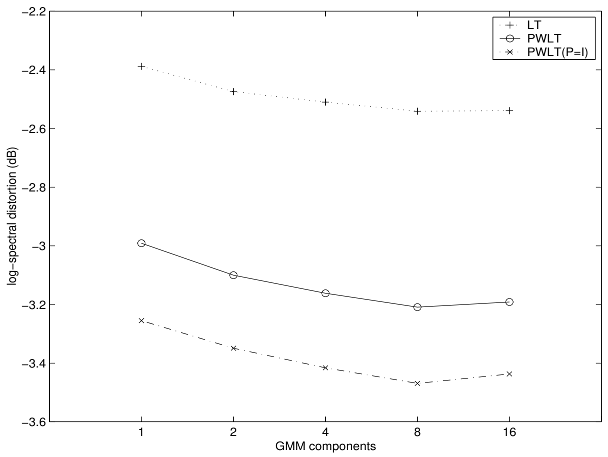


Figure 2: *Comparison between linear transformation (LT) and perceptual weighting linear transformation (PWLT) based on log-spectral distortion. Both in 25 dimension. The (P=I) case used an identical matrix as the perceptual matrix.*

where $A = [loga(\omega_1), \cdots, loga(\omega_q)]'$ is the log-spectral vector and $a(\omega)$ is the normalized amplitude at frequency $\omega$. The perceptual weight matrix $P$ is a diagonal matrix whose elements $\{p_{ii}\}$ are the perceptual weights at each frequency $\omega_i$ and it is determined by the target spectral envelope.

A popular perceptual filter in speech coding [7] and the one used here is defined by,

$$H(\omega) = \frac{A(z/\beta)}{A(z/\gamma)}, 0 < \gamma < \beta < 1 \tag{11}$$

where $A(z)$ is the LPC filter and a common choice of parameters is $\beta = 1.0$ and $\gamma = 0.8$.

Instead of using a linear transformation to transform the cepstral coefficients as in the baseline system, here the transformation function is directly applied to the spectrum. This allows the perceptual weighting to be applied directly. Using equation (3), the total perceptual spectral distance over all training data can be written as,

$$\mathcal{E} = \sum_{t=1}^{T} \Big(\mathbf{y}_t - \bar{\mathbf{W}}\Lambda(\mathbf{x}_t)\Big)' P(t) \Big(\mathbf{y}_t - \bar{\mathbf{W}}\Lambda(\mathbf{x}_t)\Big) \tag{12}$$

noting that $\mathbf{x}_t$ and $\mathbf{y}_t$ are no longer vectors of cepstral coefficients. They are now direct spectral vectors with the same form as $A_1$ and $A_2$ in equation (10).

The solution of $\bar{\mathbf{W}}$ satisfies

$$\sum_{t=1}^{T} P(t)\mathbf{y}_t\Lambda(\mathbf{x}_t)' = \sum_{t=1}^{T} P(t)\bar{\mathbf{W}}\Lambda(\mathbf{x}_t)\Lambda(\mathbf{x}_t)' \tag{13}$$

Since the left-hand side of equation (13) is independent of $\bar{\mathbf{W}}$, it will be referred to as $\mathbf{Z}$ where

$$\mathbf{Z} = \sum_{t=1}^{T} P(t)\mathbf{y}_t\Lambda(\mathbf{x}_t)' \tag{14}$$

To simplify the right-hand side of equation (13), let

$$R(t) = \Lambda(\mathbf{x}_t)\Lambda(\mathbf{x}_t)' \tag{15}$$

Equation (13) can then be rewritten as

$$\mathbf{Z} = \sum_{t=1}^{T} P(t)\bar{\mathbf{W}}R(t) \tag{16}$$

Since $P(t)$ is a diagonal matrix, a new variable $G^{(i)}$ can be defined with elements

$$g_{jq}^{(i)} = \sum_{t=1}^{T} p_{ii}(t) r_{jq}(t) \tag{17}$$

where $p_{ii}(t)$ and $r_{jq}(t)$ are the elements of $P(t)$ and $R(t)$ respectively. Then a closed form solution of $\bar{\mathbf{W}}$ can be calculated using

$$\bar{\mathbf{w}}_i = \mathbf{z}_i G^{(i)-1} \tag{18}$$

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

In practice, this closed form solution is not suitable for directly transforming the FFT spectrum since the order of the spectral vector would be too high. Fortunately, using the cubic spline interpolation technique described previously, the spectral envelope can be resampled to any number of dimensions, e.g. 25. Note however that when evaluating the objective performance the full 120 evaluation points are still used.

## 5. Comparative Evaluation

The baseline interpolated linear transformation (LT) system was first compared with the perceptually weighted linear transformation (PWLT) in terms of log-spectral distortion. Fig.2 shows the results for both cases. Also shown is the PWLT system for the case where the perceptual weighting is replaced by an identity transform (PWLT P=I). This latter case corresponds to the direct minimisation of the (unweighted) log-spectral distortion and as expected, this case results in the lowest distortion overall. Relative to this, the PWLT system has a slightly higher distortion, interestingly, the baseline LT system has a significantly higher distortion which suggests that the cepstral representation of the spectral envelope may not be an ideal choice.

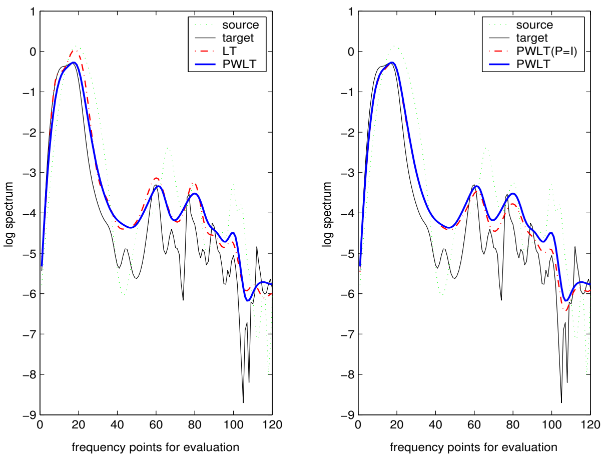


Figure 3: *Spectral envelope conversion. (a) Comparison between LT and PWLT. (b) Comparison between PWLT and PWLT(P=I). All in 25 dimension, with 8 GMM components.*

Fig.3(a) shows an example of envelope conversion using the LT and PWLT systems where it can be seen that the PWLT converted envelope has a slightly closer fit to the target envelope than the LT converted envelope. However, the differences are small. In Fig.3(b) the PWLT and the PWLT(P=I) systems are compared. The PWLT converted envelope fits more closer at the spectral peaks of the target envelope than the PWLT(P=I) envelope, but again the differences are small. Perhaps, more significantly, these examples show that significant spectral detail is lost by reducing the dimensionality of the envelope representation down to 25 coefficients. Another clear effect is the broadening of the spectral peaks caused, at least in part, by the averaging effect of least square error estimation. To evaluate the relative performance of the systems in terms of their perceptual effects, an ABX-style preference test was performed whereby 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 36 transformed utterances in total, approximately equally split between within-gender and cross-gender transformations.

Table 1 gives the percentage of the converted utterances that were labelled as closer to the target for each case. These preliminary results indicate that the performances of LT and PWLT are very similar. The PWLT scheme works slightly better than the baseline LT system, but since the P=I case is identical to the full PWLT case, it seems likely that the substantive improvement comes from the direct minimisation of the spectral envelope distortion rather than the use of a perceptual weighting. The listeners were also asked to give their preference for each pair of converted utterances. Table 2 shows that the listeners preferred 63.3% of the PWLT converted utterances when compared with the LT converted utterances. However the PWLT converted utterances did not show any quality improvement when compared with the PWLT(P=I) converted utterances. This is again consistent with the previous results

Table 1: *Results from the ABX test.*

| | LT | PWLT | PWLT(P=I) |
|---|---|---|---|
| ABX | 90.3% | 91.7% | 91.7% |

Table 2: *Results from the preference test.*

| | LT | PWLT |
|---|---|---|
| preference | 36.7% | 63.3% |

## 6. Conclusion

This paper has presented a method of voice transformation based on using perceptually weighted linear transformations. This approach applies a perceptual filter to the spectral errors and trains a linear transformation to minimize the perceptual spectral distance. When compared to a conventional system utilising a linear transform of cepstrally-encoded spectra, the proposed scheme gives slightly better performance. However, when the perceptual weighting is removed from the new scheme by substituting an identity transform, the results are unchanged suggesting that it is the direct minimisation of the log spectral distortion rather than the use of a perceptual filter which is most important.

The speech generated by all of the schemes studied had a slightly muffled effect and this is attributed partly to the averaging effect of the linear transforms and partly to the reduced dimensionality of the spectral vectors. Directly increasing the number of transforms and/or the vector size would in principle solve this problem, but as demonstrated in the paper, data sparsity problems prevent this in practice. Future work will therefore focus on methods for making more efficient use of limited training data.

## 7. Acknowledgments

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

## 8. References

[1] Abe, M., Nakamura, S., Shikano, K. and Kuwabara, H., "Voice conversion through vector quantization", in Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, 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] 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.

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

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

[7] 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.