# LOCAL LINEAR TRANSFORMATION FOR VOICE CONVERSION

*Victor Popa[1], Hanna Silen[1], Jani Nurminen[2] and Moncef Gabbouj[1]*

[1]Department of Signal Processing, Tampere University of Technology, Tampere, Finland
[2]Nokia, Tampere, Finland

## ABSTRACT

Many popular approaches to spectral conversion involve linear transformations determined for particular acoustic classes and compute the converted result as a linear combination between different local transformations in an attempt to ensure a continuous conversion. These methods often produce over-smoothed spectra and parameter tracks. The proposed method computes an individual linear transformation for every feature vector based on a small neighborhood in the acoustic space thus preserving local details. The method effectively reduces the over-smoothing by eliminating undesired contributions from acoustically remote regions. The method is evaluated in listening tests against the well-known Gaussian Mixture Model based conversion, representative of the class of methods involving linear transformations. Perceptual results indicate a clear preference for the proposed scheme.

***Index Terms***— Gaussian Mixture Model (GMM), Line Spectral Frequencies (LSF), Local Linear Transformation (LLT)

## 1. INTRODUCTION

Voice conversion is defined to be a modification of the speech signal in which the perceived speaker identity is changed while preserving the content and quality. Such a conversion involves the modification of prosodic properties as well as a transformation of the spectral features. As a core part of voice conversion, spectral transformation has received most interest in this research area and is also the scope of this article.

A remarkable number of techniques for spectral conversion have been proposed in the literature. Codebook mapping [1], GMM [2], frequency warping [3], artificial neural networks [4], linear transformations [5], bilinear models [6], eigenvoices [7] and maximum likelihood estimation of spectral parameter trajectory [8] are some of the most representative. While dealing better with one or another issue associated with the conversion all the above methods have their own limitations.

Many existing approaches involve linear conversion functions and typically suffer from two important drawbacks. One of them is related to the frame based operation in which the temporal continuity of the spectral features is ignored. The second issue is the so-called over-smoothing characterized by an undesired smoothing of the parameter tracks and converted spectra. The combined effect of these drawbacks is a poor speech quality.

The GMM based approach is very popular and representative of the class of methods based on linear transformations. In GMM based conversion, a linear transformation is trained for each Gaussian component and the result is computed as a weighted sum of local regression functions in an attempt to avoid sudden changes of the conversion function. In reality, a frame's decomposition is dominated by only one mixture component [9] making the method susceptible to discontinuities. In addition, the GMM technique is also affected by over-smoothing.

In this paper we propose a spectral conversion scheme which trains an individual linear transformation for each feature vector. The method uses an underlying codebook trained from aligned data of the two speakers and the linear transformation is computed on a selected set of codebook centers situated in the proximity of the input spectral vector in the acoustic space. By focusing on the local properties of the acoustic space, the proposed method is shown to effectively reduce the over-smoothing. Our listening tests suggest that the proposed scheme is probably affected to a lesser degree by discontinuity artifacts than the GMM approach.

While suffering serious limitations as a conversion method in itself, the codebook has the favorable property of good detail preservation which benefits the proposed algorithm where such limitations are avoided.

The article continues in Section 2 with a technical description of the proposed method. Subjective listening test results and other experiments are presented and discussed in Section 3. The article ends with conclusions and directions for future research presented in Section 4.

## 2. LOCAL LINEAR TRANSFORMATION

The use of linear transformation for spectral conversion is not new. An important number of solutions based on linear transformation have been proposed in the literature.

In [10] the aligned spectral vectors of source and target speakers are first divided into a number of classes and a linear transformation is trained for each class. All the linear transformations contribute to the conversion of each source vector in the form of a weighted sum where the weights represent probabilities that the source vector belongs to the corresponding class. The GMM based solution [2] works in a similar way using one linear transformation for each mixture component.

By analogy with [11], which argues that linear combinations over large sets of curves are bound to produce averaged results and destroy characteristic details, we believe that allowing all the linear transformations to contribute to the conversion is likely to produce a similar averaging effect equivalent to over-smoothing. Similar to Freeman et al., we believe it would be beneficial to restrict the number of linear transformations involved in conversion to only a few corresponding to the most similar speech classes. In this paper we take this idea forward and propose a local regression approach where each source vector is converted with an individual linear transformation trained locally within the neighborhood of the input vector. This method can be seen, in some sense, as a tradeoff between the mapping codebooks and, for instance, the traditional GMM approach.

Assume that our training set consists of two time aligned sequences of source and target spectral vectors, denoted $X$ and $Y$, and let us consider the codebook $M$ with $Q$ centers obtained from the quantization of sequence $Z$ of combined vectors $z_n = \left[x_n^T y_n^T\right]^T$.

$$X = [x_1, x_2, ..., x_N] \qquad Y = [y_1, y_2, ..., y_N] \tag{1}$$

$$\mathrm{M} = \begin{bmatrix} \begin{bmatrix} \mu_1^x \\ \mu_1^y \end{bmatrix} & \begin{bmatrix} \mu_2^x \\ \mu_2^y \end{bmatrix} & \dots & \begin{bmatrix} \mu_Q^x \\ \mu_Q^y \end{bmatrix} \end{bmatrix} \tag{2}$$

The idea of local regression is to fit local models to nearby data. The conversion of a source vector $x$ requires, in a first phase, the selection of a so called neighborhood of $x$ or set of codebook centers situated in the proximity of $x$. The simplest way to determine the neighborhood of $x$ is to consider its K nearest neighbors that minimize the distance:

$$d_E(x, \mu_q^x) = \left\| x - \mu_q^x \right\| \tag{3}$$

The neighborhood can be expressed formally as:

$$N(x) = \left\{ \mu_{q_1}, \mu_{q_2}, ..., \mu_{q_K} \right\} \tag{4}$$

where $q_k$ are codebook indices of the selected centers and $\mu_{q_k} = \begin{bmatrix} \mu_{q_k}^x \\ \mu_{q_k}^y \end{bmatrix}$.

In a second phase, the proposed method determines a linear transformation for each neighborhood using a least squares criterion. Local modeling favors simple models and a simple training criterion. The linear regression model is:

$$(\mu_{q_k}^x)^T \cdot W = (\mu_{q_k}^y)^T \tag{5}$$

The linear transformation W is obtained by solving:

$$N^x \cdot W = N^y \tag{6}$$

which has the least squares solution:

$$W = \left((N^x)^T N^x\right)^{-1} (N^x)^T N^y \tag{7}$$

where $N^x = \left[\mu_{q_1}^x \mu_{q_2}^x .. \mu_{q_K}^x\right]^T$ and $N^y = \left[\mu_{q_1}^y \mu_{q_2}^y .. \mu_{q_K}^y\right]^T$.

The least squares solution minimizes the criterion:

$$C = \sum_{k=1}^{K} \left\| (\mu_{q_k}^x)^T \cdot W - (\mu_{q_k}^y)^T \right\|^2 \tag{8}$$

Finally, the converted result for $x$ is computed as:

$$(y_{conv})^T = x^T \cdot W \tag{9}$$

The conversion of an entire sequence of source vectors can be obtained by repeating for each vector the procedure described above.

In practice it was noticed that the quality of the conversion is sensitive to the selected neighborhood and the type of linear transformation used. Firstly, it was found to be beneficial to estimate band diagonal matrices instead of full ones given that the correlation is highest between neighbor elements of an LSF vector. Secondly, it was found beneficial to use $y_{conv}$ for a new selection of neighbors minimizing:

$$d_E\left(\begin{bmatrix} x \\ y_{conv} \end{bmatrix}, \mu_q\right) = \left\| \begin{bmatrix} x \\ y_{conv} \end{bmatrix} - \mu_q \right\| \tag{10}$$

and iterate the same steps until the neighborhoods determined in consecutive steps become virtually identical or sufficiently similar. This is equivalent to a convergence of $y_{conv}$. The process was found to be pseudo-convergent and can be stopped with an arbitrary threshold criterion. Figure 1 illustrates this pseudo-convergence.

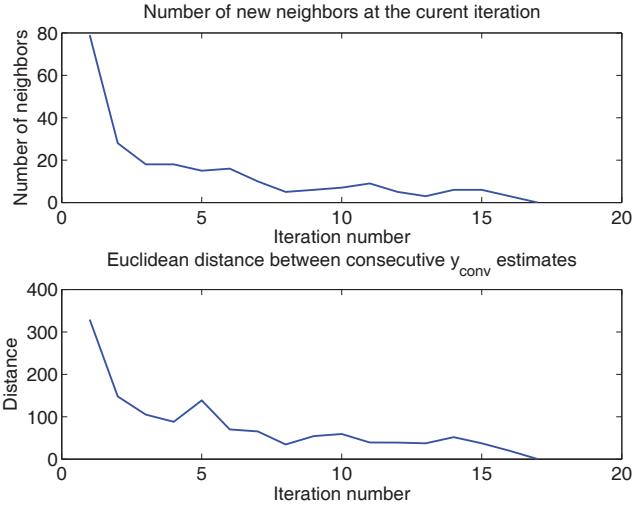


**Figure 1. Pseudo-convergence of neighborhood selection**

We observe that the algorithm could have been applied directly on the aligned training data instead of the codebook.

## 3. EXPERIMENTS

The algorithm for spectral conversion presented in the previous section has been applied on 16-dimensional line spectral frequencies (LSF) vectors and the results are demonstrated in this section with two cross gender examples. The section presents a comparison with the popular GMM based approach providing objective and subjective results.

### 3.1. Acoustic Data

CMU Arctic database (http://festvox.org/cmu_arctic/) is a publicly available corpus of parallel speech sampled at 16 kHz. We used the CLB (female) and RMS (male) speakers from the CMU Arctic database to test conversion in both directions: from male voice to female and from female voice to male.

A parallel set of 100 sentences was used as training data amounting to approximately 30000 pairs of source and target LSF vectors after time alignment. Another 10 sentences were used for testing.

### 3.2. Model Settings

*3.2.1. GMM*

Too few components, although reliably estimated, give an inaccurate approximation of the training data while the estimation of too many components is unreliable causing over-fitting. In choosing a reference GMM for the comparison with the proposed approach such problems are avoided as follows. The performance of GMM models with different numbers of components was evaluated over the test set and the model with the lowest error was selected.

As illustrated in Figure 2 the female to male direction requires 8 components while 16 components are needed to convert the male into female voice. The mean squared error (MSE) figures are based on the definition given in [6].

Even though the GMM was tuned directly on the test set, a similar tuning could be performed by cross-validation using only the training set.

*3.2.2. Proposed Method (Local Linear Transformation)*

The tuning of the proposed method is mainly based on perceptual evaluation. A codebook size of 8000 was used while the neighborhood sizes were tuned separately for each direction leading to values of 40 (female to male) and 130 (male to female). The linear transformations were restricted to tri-diagonal matrices.

The neighborhood size was found to act as a tradeoff producing unstable results when the neighborhood is too small and excessively averaged (over-smoothed) results when neighborhoods are large.

### 3.3. Subjective Listening Test

The speech samples evaluated in the listening tests are based on target speaker versions of the test utterances, in whose parametric representations only LSFs have been replaced with converted ones. This mimics the case when all other features are ideally converted focusing the evaluation on the actual spectral conversion.

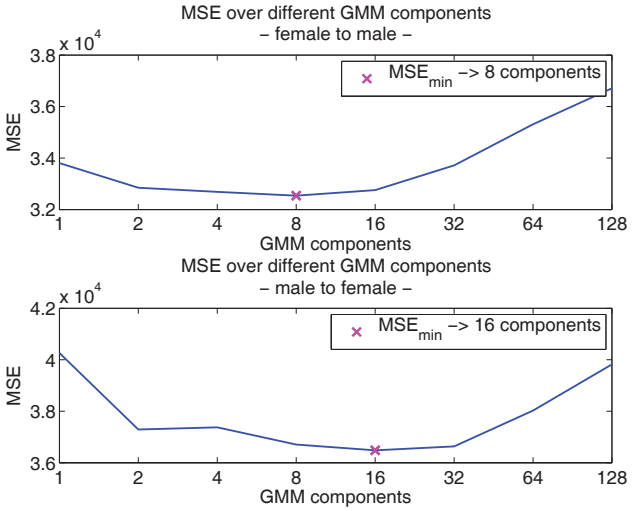


**Figure 2. Mean squared error of GMMs for different numbers of components measured over the test set**

For each conversion direction a modified MOS test was carried out by ten listeners on ten test sentences. The proposed method (LLT) and the GMM based approach were compared in terms of speech quality and success of identity mapping. These criteria are evaluated with scores between -2 and 2 with -2/+2 indicating that "GMM/LLT performs much better", -1/+1 for "GMM/LLT performs better" and 0 indicating perceptually identical performance. The results of the listening test are illustrated in Table 1.

**Table 1. Subjective listening test scores with 95% confidence intervals.**

| | Quality | Identity |
|---|---|---|
| Female to male | 0.49±0.19 | 0.33±0.17 |
| Male to female | 0.23±0.17 | 0.15±0.16 |

A possible explanation for the male to female result is that the high pitched female voice seems to mask the quality problems making the two methods sound more similar.

The subjective scores indicate the general preference of the proposed approach over the GMM based system.

### 3.4. Over-Smoothing Reduction

The converted spectra and LSF tracks illustrated in Figure 4 indicate a reduction of the over-smoothing in the case of the proposed approach in comparison to GMM.

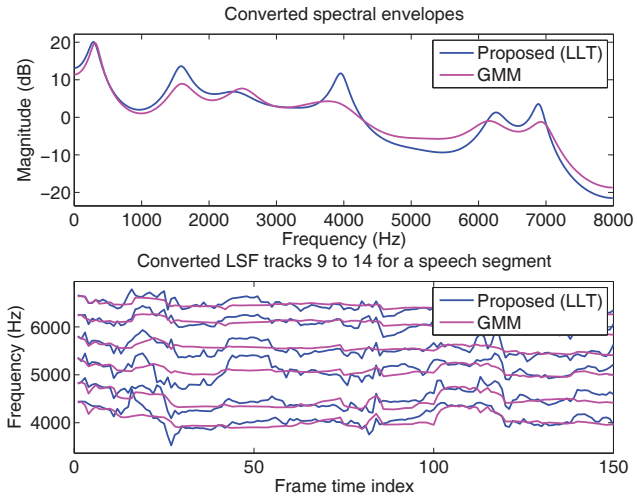


**Figure 3. Over-smoothing reduction for spectral envelopes (top) and LSF tracks (bottom).**

Standard deviation measurements of converted and original target spectra (in frequency) and LSF tracks (in time) are calculated over the entire test set and summarized in Table 2 confirming the over-smoothing reduction.

**Table 2. Average standard deviation of spectral magnitude (in dB) and LSF tracks (in Hz)**

| | Magnitude (dB) | | | LSF tracks (Hz) | | |
|---|---|---|---|---|---|---|
| | Proposed | GMM | Tgt | Proposed | GMM | Tgt |
| Female to male | 8.19 | 7.46 | 8.74 | 237 | 199 | 264 |
| Male to female | 10.70 | 9.86 | 10.65 | 336 | 296 | 328 |

Local modeling, rather than the interpolation of local models from acoustically remote regions, makes the proposed approach capable to capture details better and reduce the averaging effect.

## 4. CONCLUSIONS

This article introduced a new method for spectral conversion based on local regression. Linear transformation models are fit every time to local data for each source vector as opposed to the typical interpolation of linear models. The method was shown to effectively reduce over-smoothing and obtained favorable preference scores in a subjective evaluation against the popular GMM based approach.

On the downside the proposed method uses heavier computation for conversion as linear transformations depend on the input vector and have to be estimated at runtime.

Interesting directions for future work would be to study alternative ways for neighborhood selection and alternative local models.

## 5. ACKNOWLEDGEMENT

This work was supported by the Academy of Finland, (application number 129657, Finnish Programme for Centres of Excellence in Research 2006-2011).

## 6. REFERENCES

[1] L. Arslan, and D. Talkin, "Voice Conversion by Codebook Mapping of Line Spectral Frequencies and Excitation Spectrum," In 5th *Proceedings of European Conference on Speech Communication and Technology*, Rhodes, Greece, 1997.

[2] A. Kain and M.W. Macon, "Spectral Voice Conversion for Text-to-Speech Synthesis," In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, Seattle, USA, Vol.1, pp. 285-288, 1998.

[3] Z. Shuang, R. Bakis and Y. Qin, "Voice Conversion Based on Mapping Formants," In *Proceedings of TC-Star Workshop on Speech to Speech Translation*, Barcelona, Spain, June 2006, pp. 219-223.

[4] M. Narendranath, H. Murthy, S. Rajendran, and N. Yegnanarayana, "Transformation of Formants for Voice Conversion Using Artificial Neural Networks", *Speech Communication*, vol.16, pp. 207-216, 1995.

[5] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion", *IEEE Transaction on Speech and Audio Processing*, vol. 6, no.2, pp.131-142, 1998.

[6] V. Popa, J. Nurminen, and M. Gabbouj, "A Study of Bilinear Models in Voice Conversion," *Journal of Signal and Information Processing*, vol. 2, no.2, May 2011.

[7] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice Conversion Based on Gaussian Mixture Model", *Proc. ICSLP*, pp. 2446-2449, Pittsburgh, USA, Sep. 2006.

[8] T. Toda, A.W. Black, and K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory", *IEEE Transactions on Audio, Speech and Language Processing*, Volume 15, Issue 8, pp. 2222-2235, Nov. 2007.

[9] E. Helander, T. Virtanen, J. Nurminen and M. Gabbouj, "Voice Conversion Using Partial Least Squares Regression," *IEEE Trans. on Speech and Audio Processing*, vol. 18, no. 5, pp. 912-921, July 2010.

[10] H. Ye and S. Young, "Perceptually Weighted Linear Transformations for Voice Conversion," *Eurospeech*, Geneva, Switzerland, 2003.

[11] W.T. Freeman, J.B. Tenenbaum and E. Pasztor, "Learning Style Translation for the Lines of a Drawing," *ACM Transactions on Graphics*, vol. 22, no. 1, pp. 33-46, Jan. 2003.