Voice Conversion using GMM with Enhanced Global Variance

Hadas Benisty and David Malah

Department of Electrical Engineering
Technion, Israel Institute of Technology
Haifa, 32000, Israel
{hadasbe@tx,malah@ee}.technion.ac.il

Abstract

The goal of voice conversion is to transform a sentence said by one speaker, to sound as if another speaker had said it. The classical conversion based on a Gaussian Mixture Model and several other schemes suggested since, produce muffled sounding outputs, due to excessive smoothing of the spectral envelopes.

To reduce the muffling effect, enhancement of the Global Variance (GV) of the spectral features was recently suggested. We propose a different approach for GV enhancement, based on the classical conversion formalized as a GV-constrained minimization. Listening tests show that an improvement in quality is achieved by the proposed approach.

Index Terms: Voice conversion, GMM, Global Variance (GV)

1. Introduction

A voice conversion system aims to convert the perceived identity of a sentence said by a source speaker to that of a given target speaker. Such transformation is useful in applications that synthesize speech signals based on prerecorded sentences, such as Text-To-Speech (TTS) systems and automatic dialog systems.

One of the earliest spectral envelope conversion methods, presented in [1], proposed a codebook based conversion. This method used hard clustering and mapping, but the converted output suffered from poor quality due to coarse quantization.

A more flexible approach was proposed in [2] and is the most commonly used voice conversion method to date. It uses a Gaussian Mixture Model (GMM) to statistically model the spectral features of the source speaker. The conversion function is interpreted as the expected value of the target spectral envelope, given the source spectral envelope, and therefore takes the form of a linear estimator. The conversion parameters are evaluated by Least Squares (LS) using a parallel training set. In [3] a joint source-target GMM training was proposed, where the conversion function takes the same linear form as in [2], but its parameters are evaluated during the joint-GMM training process. Due to its statistical modeling and linear conversion, the converted spectra using those approaches are overly smoothed, leading to a muffled output signal. Several modifications of this method have been proposed since, among these: GMM & Dynamic Frequency Warping (DFW), [4], GMM & codebook selection, [5] and a combined pitch & spectral envelope GMM-based conversion, [6]. Still, these GMM-based conversion methods report to produce a muffled output signal, probably due to excessive smoothing of the temporal evolution of the spectral envelope. Another approach, presented in [7], aims to capture the temporal evolution of the spectral envelope, and to increase the global variance (GV) of the spectral features. Sequences of spectral features and their variances are jointly modeled using GMM, and the conversion is evaluated using a Maximum Likelihood (ML) estimation.

In this paper we propose a different approach for GV enhancement using the classical conversion proposed in [2]. We formalize the training process as a constrained least squares minimization problem: the mean distance between the converted and target features is minimized under the constraint that the GV of the converted features should match the GV of the target features. Objective tests show that compared to the classical method, the proposed approach increases the GV of the spectral features, but the spectral similarity to the target is somewhat reduced. Nevertheless, subjective evaluations indicate that the output of the constrained conversion is preferable in terms of both quality and similarity to the target.

The organization of this paper is as follows. In Sec. 2, the classical GMM-based conversion, using LS estimation, is described. The proposed GV enhancement approach is presented in Sec. 3. Objective and subjective evaluation results are presented in Sec. 4. Conclusions and suggestions for further work are summarized in Sec. 5.

2. Classical Voice Conversion using GMM and Least Squares (LS)

Let $\{\mathbf{x^q}\}_{q=1}^{Q} \in \mathbf{R}^P$ be a set of feature vectors, representing the spectral characteristics of a set of sentences said by a source speaker, and $\{\mathbf{y^q}\}_{q=1}^{Q} \in \mathbf{R}^P$ a similar set corresponding to a target speaker. The two sets are assumed to originate from a parallel, time aligned, training set. The classical statistical voice conversion method, [2], uses a Gaussian Mixture Model (GMM) as a statistical model for the feature vectors related to the source speaker. The source vectors are divided into M classes, and the vectors in every class are assumed to be jointly Gaussian, so that the probability of a source vector $\mathbf{x}^q$ is:

$$\begin{aligned} p\left(\mathbf{x}^q\right) &= \sum_{m=1}^{M} p\left(w_m\right) N\left(\mathbf{x}^q; \mu^{(x),m}, \mathbf{\Sigma}^{(xx),m}\right) \\ q &= 1, ..., Q \end{aligned} \tag{1}$$

where $p\left(w_m\right)$ is the probability of class w_m, and $N\left(\cdot; \mu^{(x),m}, \mathbf{\Sigma}^{(xx),m}\right)$ is a normal distribution, with mean vector $\mu^{(x),m}$, and covariance matrix $\mathbf{\Sigma}^{(xx),m}$. The parameters of the GMM are usually estimated using the Expectation Maximization (EM) [8] algorithm fed with the training vectors related to the source speaker. The conversion

function proposed by [2] has a linear form:

$$\mathcal{F}\{\mathbf{x}\} = \sum_{m=1}^{M} p\left(w_m|\mathbf{x}\right)\left(\nu^m + \right. \\ \left. + \ \mathbf{\Gamma}^m(\mathbf{\Sigma}^{(xx),m})^{-1}\left(\mathbf{x}-\mu^{(x),m}\right)\right), \quad (2)$$

where $p\left(w_m|\mathbf{x}\right)$ is a conditional probability evaluated using the GMM parameters and Bayes' theorem:

$$p\left(w_m|\mathbf{x}\right) = \frac{p\left(w_m\right)N\left(\mathbf{x};\mu^{(x),m},\mathbf{\Sigma}^{(xx),m}\right)}{\sum_{m=1}^{M}p\left(w_m\right)N\left(\mathbf{x};\mu^{(x),m},\mathbf{\Sigma}^{(xx),m}\right)} \quad (3)$$

and $\{\mathbf{\Gamma}^m,\nu^m\}_{m=1}^M$ are the $P\times P$ and $P\times 1$ conversion matrices and vectors. These parameters are evaluated so that the mean Euclidian distance between the converted and target spectral features is minimized:

$$\min_{\{\mathbf{\Gamma}^m,\nu^m\}_{m=1}^M}\frac{1}{Q}\sum_{q=1}^{Q}\|\mathcal{F}\{\mathbf{x}^q\}-\mathbf{y}^q\|^2. \quad (4)$$

Training a full covariance GMM requires a large training set. It is commonly assumed that the spectral features are statistically independent, so that the covariance matrices $\mathbf{\Sigma}^{(xx),m}$ are diagonal. In this case the training process requires a much smaller training set. In addition, the matrices $\{\mathbf{\Gamma}^m\}_{m=1}^M$ are also diagonal, so their estimation in (4) can be separated into P independent minimization problems - one for every coordinate, $p=1,...,P$:

$$\min_{\{\gamma_p^m,\nu_p^m\}_{m=1}^M}\frac{1}{Q}\sum_{q=1}^{Q}\|\mathcal{F}\{x_p^q\}-y_p^q\|^2 \quad (5)$$

where $\{\gamma_p^m\}_{m=1}^M$ are the (p,p) elements of $\{\mathbf{\Gamma}^m\}_{m=1}^M$, and $\{\nu_p^m\}_{m=1}^M$ are the p-th elements of $\{\nu^m\}_{m=1}^M$.

As described in [2], a matrix form of (5) can be formulated as:

$$\min_{\mathbf{q}^p}\|\mathbf{A}^p\mathbf{q}^p-\mathbf{y}_p\|^2, \quad (6)$$

where $\mathbf{y}_p$ is a $Q\times 1$ vector including the p-th element of all the target training vectors, defined in (7), $\mathbf{A}^p$ is a $Q\times 2M$ matrix defined in (8), and $\mathbf{q}^p$ is a $2M\times 1$ vector including the conversion parameters defined in (9).

$$\mathbf{y}_p \triangleq \left(\begin{array}{ccc} y_p^1 & \cdots & y_p^Q \end{array}\right)^T, \quad (7)$$

$$\begin{aligned} \mathbf{A}^p &\triangleq \left(\begin{array}{ccc} \mathbf{P} & \vdots & \mathbf{D}^p \end{array}\right) \\ \{\mathbf{P}\}_{m,q} &= p\left(w_m|\mathbf{x}^q\right) \\ \{\mathbf{D}^p\}_{q,m} &= p\left(w_m|\mathbf{x}^q\right)\frac{1}{\sigma^{q,m}}\left(\mathbf{x}_p^q-\mu_p^{(x),m}\right) \\ m &= 1,...,M;\quad q=1,...,Q \end{aligned} \quad (8)$$

$$\mathbf{q}^p \triangleq \left(\begin{array}{ccccccc} \nu_p^1 & \cdots & \nu_p^M & \vdots & \gamma_p^1 & \cdots & \gamma_p^M \end{array}\right)^T. \quad (9)$$

The LS solution of (6) is given by:

$$\hat{\mathbf{q}}^p = \left(\mathbf{A}^{pT}\mathbf{A}^p\right)^{-1}\mathbf{A}^{pT}\mathbf{y}_p \quad (10)$$

In the next section we propose a new approach for GV enhancement, in the framework of the classical GMM-based conversion, using a constrained LS minimization.

3. GMM-based Conversion with a GV Constraint

The GV of the p-th element of the target feature vectors can be evaluated by:

$$Var\{\mathbf{y}_p\} \simeq \frac{1}{Q}\sum_{q=1}^{Q}\left(y_p^q-\frac{1}{Q}\sum_{r=1}^{Q}y_p^r\right)^2 \quad (11)$$

where $\mathbf{y}_p$ is the $Q\times 1$ vector defined in (7). A matrix form of the r.h.s. of (11) is:

$$\frac{1}{Q}\sum_{q=1}^{Q}\left(y_p^q-\frac{1}{Q}\sum_{r=1}^{Q}y_p^r\right)^2 = \|\mathbf{\Delta}\cdot\mathbf{y}_p\|^2 \triangleq c_p^2, \quad (12)$$

where $\mathbf{\Delta}$ is a $Q\times Q$ matrix defined by:

$$\mathbf{\Delta} \triangleq \frac{1}{\sqrt{Q}}\left(\mathbf{I}_{Q\times Q}-\frac{1}{Q}\left(\begin{array}{ccc} 1 & \cdots & 1 \\ \vdots & \ldots & \vdots \\ 1 & \cdots & 1 \end{array}\right)\right) \quad (13)$$

Similarly, the GV of the p-th element of the converted vectors can be evaluated by:

$$Var\{\mathcal{F}\{\mathbf{x}_p^q\}\} \simeq \|\mathbf{\Delta}\cdot\mathbf{A}^p\mathbf{q}^p\|^2 = \|\mathbf{B}^p\mathbf{q}^p\|^2, \quad (14)$$

where $\mathbf{B}^p \triangleq \mathbf{\Delta}\cdot\mathbf{A}^p$.

In order to enhance the GV of the converted elements, while minimizing the mean Euclidian distance between the converted and target vectors, we propose a constrained formulation:

$$\begin{array}{l} \min_{\mathbf{q}^p}\|\mathbf{A}^p\mathbf{q}^p-\mathbf{y}_p\|^2 \\ s.t.\ \ \|\mathbf{B}^p\mathbf{q}^p\|^2 = c_p^2 \end{array} \quad ;\quad p=1,...,P \quad (15)$$

where c_p^2 is the evaluated GV of the target, defined in (12).

The P constrained minimization problems defined in (15) can be solved by using the Lagrange Multiplier method and joint diagonalization of the pairs $\{\mathbf{A}^p,\mathbf{B}^p\}_{p=1}^P$, as described in [9].

4. Experimental Results

4.1. Experimental Conditions

We used two U.S. English male speakers from the CMU ARCTIC database [10]: 50 parallel sentences were used for training and 50 other parallel sentences for testing, all sampled at 16kHz and phonetically annotated. Analysis and synthesis were performed using the Harmonic Plus Noise Model (HNM) [11] by the toolkit available at [12]. The first 24 Mel Frequency Cepstrum Coefficients (MFCC's) were extracted using the harmonic amplitudes as described in [13]. The analysis frames were time aligned and the feature vectors were matched using a DTW algorithm based on the phonetic labeling as described in [14].

The dynamic range of the cepstral coefficients in natural speech usually decreases as their order increases, so enhancement of the high order coefficients in the converted signal is not as important as for low order coefficients. Therefore, the GV was enhanced only for cepstral coefficients lower than a specific threshold $P_0=12$. For $p>P_0$ the conversion parameters were evaluated using the classical, unconstrained approach.

The pitch was converted linearly, so that the mean and standard deviation of the converted pitch will match the mean and variance of the pitch values related to the target speaker:

$$\hat{f}_0^{(y)}(k) = \mu^{(y)} + \frac{\sigma^{(y)}}{\sigma^{(x)}}\left(f_0^{(x)}(k)-\mu^{(x)}\right), \quad (16)$$

where $f_0^{(x)}(k)$ and $\hat{f}_0^{(y)}(k)$ are the pitch values of the source and converted signals at the k-th frame, respectively. The parameters $\mu^{(x)}$ and $\mu^{(y)}$ are the mean pitch values and $\sigma^{(x)}$ and $\sigma^{(y)}$ are the standard deviations of the source and target pitch values, respectively.

The performance of our proposed conversion method was examined and compared to the performance of classical conversion [2], using both objective and subjective measures. To reduce audible artifacts converted outcomes by both methods were processed. Before synthesis, the temporal evolution of each cepstral coefficient was filtered by $\mu + (1-\mu)\, z^{-1}$, using $\mu = 0.5$. After synthesis, the waveforms were filtered using a low-pass filter having a 5kHz cut-off frequency.

4.2. Objective Evaluations

The similarity of the converted signals to the target signals was evaluated using mean Log Spectral Distortion (LSD). MFCC's were used as spectral features, so the LSD between the converted and target spectral envelope at each frame can be evaluated using the Euclidean distance between their corresponding feature vectors, $\mathcal{F}\{\mathbf{x}\}$ and $\mathbf{y}$:

$$LSD\left(\mathcal{F}\{\mathbf{x}\}, \mathbf{y}\right) \approx \frac{10}{ln10}\sqrt{2\sum_{p=1}^{P}\left\|\mathcal{F}\{x_p\} - y_p\right\|^2}, \quad (17)$$

where $\mathcal{F}\{x_p\}$ and y_p are the p-th elements of the source and target MFCC vectors ,respectively, and P is the length of the cepstral feature vectors.

The proposed and classical methods were also examined in terms of mean normalized GV. The GV of each converted cepstral coefficient was normalized by its corresponding natural value related to the target speaker. The first P_0 normalized GV's were averaged to produce:

$$\text{Mean Norm. GV} = \frac{1}{P_0}\sum_{p=1}^{P_0}\frac{\text{Converted GV}\,(p)}{\text{Target GV}\,(p)} \quad (18)$$

Table 1: *Objective performance of the proposed approach (labeled as Constrained GMM) compared to the Classical GMM-based method.*

Conversion Method	Mean LSD [dB]	Mean Norm. GV
Classical GMM	6.2	0.1
Constrained GMM	7.3	0.9

As seen in Table 1, the proposed approach increased the mean normalized GV from 10% to 90% of its natural value, at the expense of a degradation of $1.1dB$, in the LSD. The mean normalized GV achieved by the constrained approach did not reach 100%, though it was constrained to match the natural value of the target signal, since the test sentences were not included in the training set.

Subjective listening tests, presented in the next subsection, demonstrate the improved quality of the proposed method, compared to the classical method.

4.3. Subjective Evaluations

Listening tests were used to evaluate the performance of the proposed constrained approach, compared to the classical method, in terms of quality and individuality. We conducted two quality tests: a preference test and a Multi Stimulus test with Hidden Reference and Anchor (MUSHRA) [15], and one individuality XAB test (as conducted in [7]). In every test, 10 different sentences were examined by 10 listeners (voice samples can be listened to via the link in [16]). The group of listeners included 20-30 years old, non-experts, men and women.

In the quality preference test the listeners were asked to indicate which sentence is of better quality, when presented with two randomly ordered, converted signals. The examined signals were outputs of the proposed constrained method and outputs of the classical conversion method. The results, presented in Fig. 1, indicate that the enhanced output was almost always (about 95% of the time) preferred by the listeners.

The purpose of the MUSHRA test, [15], is to evaluate the quality of several processed signals compared to a given high quality, usually unprocessed, reference signal. The listeners are presented with several test signals including: outputs of the examined systems, a hidden anchor signal - usually a filtered version of the reference signal, and a hidden reference. The test signals are randomly ordered, and the listeners are not informed about the hidden reference and anchor signals being included in the test set. During evaluation, the listeners are asked to compare the reference signal to the test signals and rate them between 0 to 100, where at least one of the signals must be rated 100.

We conducted MUSHRA tests to evaluate the perceived quality of the converted outputs, compared to the quality of the original target signal. The original 10 (unprocessed) target sentences were used as reference signals. Four versions were presented as test samples: (1) A converted outcome by the proposed method. (2) A converted outcome by the classical method. (3) A hidden anchor - the target signal, low-pass filtered with a 3.5kHz cut-off frequency. (4) A hidden reference - the original unprocessed target signal. All of the listeners rated the hidden target signal as 100, and the anchor received a mean score of 80. The grades of the converted outputs presented in Fig. 2, demonstrate the improved quality achieved by the proposed constrained approach, compared to the classical approach.

In the XAB individuality test, the listeners were presented with two converted outputs (by the proposed and classical methods), randomly marked as A or B, and a processed version of the target signal, marked as X. The target signal was processed by the same tools used for processing the converted outputs. First, the target waveform was analyzed, and its cepstral coefficients were filtered by $\mu + (1-\mu)\, z^{-1}$, using $\mu = 0.5$. Then the waveform was re-synthesized and filtered using a low-pass filter having a 5kHz cut-off frequency. The results of the individuality preference test are presented in Fig. 3. They indicate that in 75% of the tests the enhanced outputs were perceived as more similar to the target signal than those obtained by the classical method, even though the constrained approach suffers from some degradation in terms of mean spectral distance.

5. Conclusion

The classical spectral envelope conversion approach is based on GMM modeling and linear conversion. This method and several others that were suggested since, are reported to suffer from a muffling effect, ascribed to excessive smoothing of the spectral envelopes. An existing method based on ML estimation [7], deals with the muffling effect by increasing the GV of the spectral features.

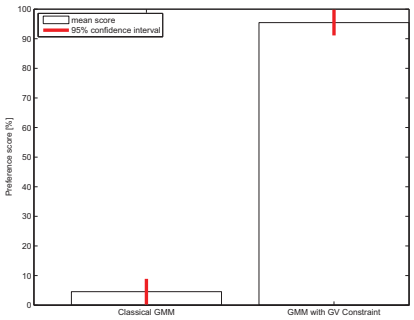

Figure 1: *Preference quality test - the classical GMM conversion against the proposed constrained conversion.*

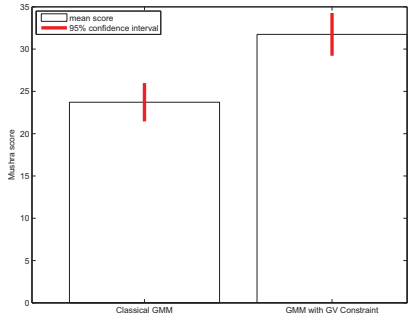

Figure 2: *MUSHRA quality test - the classical GMM conversion and the proposed constrained conversion.*

In this paper, we propose a different method for GV enhancement based on the classical conversion method. The training process is formalized as a constrained LS problem. The spectral distance between the converted and target signals is minimized, under a constraint that the GV of the converted features should match the GV of the target sentences.

Experimental results show that the proposed approach significantly increased the GV of the converted spectral features. However, the mean spectral distortion obtained by the proposed approach is somewhat higher than the mean distance achieved by the classical approach. Still, subjective evaluations indicate that the signals obtained by the proposed approach are mostly preferred by the listeners in terms of both quality and similarity to the target speaker, when compared to the converted outputs of the classical method.

Further work can be done regarding GV enhancement in case of non-diagonal covariance GMM. As opposed to the diagonal case presented above, when using full-covariance matrices the constrained minimization is not separable. Therefore, the main challenge is to overcome the computational complexity involved in the full covariance case.

6. Acknowledgements

The authors would like to thank R. Hoory, the head of the Speech Technologies Group in IBM, Haifa Research Laboratory (HRL), S. Shechtman, and Z. Kons (both from HRL), for their support and useful discussions in the course of the work.

7. References

[1] M. R. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Acoustics, Speech, and Signal Processing, 1998. Proceedings. (ICASSP '98). IEEE International Conference on*, 1988, pp. 655–658.

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

[3] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Acoustics, Speech, and Signal Processing, 1998. Proceedings. (ICASSP '98). IEEE International Conference on*, 1998, pp. 285–288.

[4] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of straight spectrum," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). IEEE International Conference on*, 2001, pp. 841–844.

[5] A. Kain and M. W. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). IEEE International Conference on*, 2001, pp. 813–816.

[6] T. En-Najjary, O. Rosec, and T. Chonavel, "A voice conversion method based on joint pitch and spectral envelope transformation," in *Proceedings of Interspeech ICSLP*, 2004, pp. 1225–1225.

[7] T. B. A. Toda and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 15, no. 8, pp. 2222–2235, 2007.

[8] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. R. Stat. Soc. B*, vol. 39, pp. 1–38, 1977.

[9] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. The Johns Hopkins University Press, 1996.

[10] J. Kominek and A. W. Black, "CMU ARCTIC databases for speech synthesis," 2003.

[11] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. on Speech and Audio Proc.*, vol. 9, no. 1, pp. 21–28, 2001.

[12] D. Erro and A. Moreno, "Online: http://www.talp.cat/talp/index.php/ca/recursos/eines/voice-conversion."

[13] O. Cappe and E. Moulines, "Regularization techniques for discrete cepstrum estimation," *IEEE Signal Processing Letters*, vol. 3, no. 4, pp. 100–102, 1996.

[14] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 18, no. 5, pp. 922–931, 2010.

[15] "ITU-R Recommendation, Method for the subjective assessment of intermediate sound quality (MUSHRA)," in *International Telecommunication Union, BS. 1534-1*, 2001, pp. BS. 1534–1.

[16] http://sipl.technion.ac.il/Info/hadas/sound-samples.htm

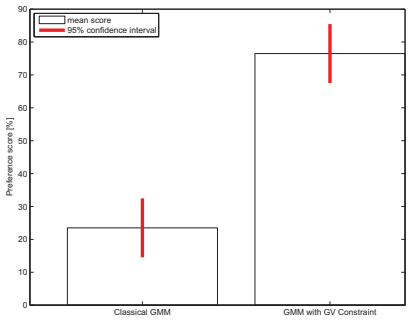

Figure 3: *Preference individuality test - the classical GMM conversion against the proposed constrained conversion.*