MODULAR GLOBAL VARIANCE ENHANCEMENT FOR VOICE CONVERSION SYSTEMS

H. Benisty, D. Malah, and K. Crammer

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

ABSTRACT

Voice conversion systems aim to transform sentences said by one speaker, to sound as if another speaker had said them. Many statistically trained conversion methods produce muffled synthesized outputs due to over-smoothing of the converted spectra. To deal with the muffling effect, conversion methods integrated with Global Variance (GV) enhancement, have been proposed. In order to gain the benefits of GV enhancement, the user is restricted to apply one of these methods as a conversion method.

We propose a new GV enhancement method designed independently of any specific conversion scheme and applied as a post-processing block. The extent of GV enhancement is controlled through the allowed spectral distance between the enhanced and the originally converted output, as specified by the user. Listening tests showed that the proposed method improves both quality and similarity to the target of the examined converted sentences, outperforming other enhancement approaches that we evaluated.

Index Terms— Global Variance (GV), Gaussian Mixture Model (GMM), Log Spectral Distance (LSD), Voice Conversion.

1. INTRODUCTION

The goal of a voice conversion process is to modify spoken sentences such that the perceived speaker is changed. Voice conversion systems are especially useful for personalizing the output of Text-to-Speech (TTS) systems, and can also be applicable for voice restoration, in case of vocal pathology, or for entertainment purposes.

The identity of a speaker is mainly associated with the spectral envelope of the speech signal, and its prosody attributes: pitch, duration, and energy. Most voice conversion methods address the spectral envelope, while the prosody values are usually linearly adjusted to match the mean values of the features associated with the target speaker.

One of the first spectral envelope conversion methods was based on codebook selection using hard clustering and mapping [1]. The resulting converted speech suffered from poor quality due to coarse quantization.

Later, a more flexible approach was proposed using a Gaussian Mixture Model (GMM) as a statistical tool for characterizing the spectral envelope of the source speaker. Least Squares (LS) approximation was used to obtain a linear conversion between the source and target envelopes [2].

In another approach [3], the conversion parameters were estimated by training a joint GMM for the source and target feature vectors. Due to the averaging process used in statistical modeling, these GMM-based methods produce overly smoothed spectral envelopes, leading to muffled synthesized outputs. Still, these are two of the most popular approaches for spectral voice conversion to date.

Several modifications of the GMM-based conversion have been proposed since, among these: GMM & codebook selection [4], GMM & Dynamic Frequency Warping (DFW) [5], GMM & Weighted Frequency Warping [6], and GMM & partial least squares regression [7]. Yet, these GMM-based conversion methods report to produce muffled output speech, apparently due to excessive smoothing of the temporal evolution of the spectral envelope.

Another approach [8] aims to capture the temporal evolution of the spectral envelope by applying Maximum Likelihood (ML) estimation. This approach also enhances the Global Variance (GV) of the spectral features, thus increasing their dynamic range, and hence decreasing the muffling effect. A different approach for GV enhancement was proposed recently [9]. It uses the framework of the classical GMM training, while constraining the GV of the converted feature vectors to match its evaluated value for the target speaker. Thus, the GV enhancement performed by these two methods is intergraded into the conversion process.

We propose a new method for GV enhancement, designed independently of a specific conversion procedure. Given a sequence of converted feature vectors, their enhanced version is obtained by maximizing their GV, under a spectral distortion constraint. The GV of the enhanced sequences is increased up to the level where the mean spectral distance between the converted sequence and its enhanced version reaches a pre-

set threshold value. The method described here enables the user to adjust the degree of GV enhancement by setting the threshold for the allowed spectral distance from the originally converted signal.

We evaluate our GV enhancement method by applying it as a post-processing block on converted outcomes of the classical GMM method [2]. The enhanced sentences were compared to the original converted sentences, and also to sentences converted (with integrated enhancement) by the Constrained GMM (CGMM) method [9]. Listening tests showed that the proposed GV enhancement method improved the quality of sentences converted by the classical GMM method [2]. In addition, most listeners preferred these results over converted sentences obtained by CGMM [9], both in terms of quality and similarity to the target speaker.

2. CLASSICAL VOICE CONVERSION USING GMM

We begin by briefly describing the classical GMM-based voice conversion method [2]. Let $\{\mathbf{x^q}\}_{q=1}^{Q}, \{\mathbf{y^q}\}_{q=1}^{Q} \in \mathrm{R}^P$ be a parallel and aligned training set consisting of feature vectors related to the source and target speakers, respectively. The source vectors are modeled using a GMM distribution. The probability of a source feature vector $\mathbf{x}^q$ is:

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

where M is the number of Gaussian components, $p\left(w_m\right)$ is the probability of component w_m and $\mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}^m, \boldsymbol{\Sigma}^m\right)$ is a normal distribution, with mean vector $\boldsymbol{\mu}^m$, and covariance matrix $\boldsymbol{\Sigma}^m$. The Expectation Maximization (EM) algorithm is commonly used to estimate the parameters of the GMM distribution based on the source training set.

Using the trained GMM parameters a linear conversion function is formed:

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

where the conditional probability $p\left(w_m|\mathbf{x}\right)$ is evaluated using the GMM parameters and Bayes' theorem. The conversion parameters $\{\boldsymbol{\Gamma}^m, \boldsymbol{\nu}^m\}_{m=1}^{M}$ are $P \times P$ and $P \times 1$ matrices, evaluated so that the mean squared-error between the converted and target spectral features is minimized:

$$\min_{\{\boldsymbol{\Gamma}^m, \boldsymbol{\nu}^m\}_{m=1}^{M}} \frac{1}{Q} \sum_{q=1}^{Q} \|\mathcal{F}\{\mathbf{x}^q\} - \mathbf{y}^q\|_2^2, \tag{3}$$

where $\|\cdot\|_2$ denotes ℓ_2 norm.

Next, we describe our proposed stand-alone method for GV enhancement. This method provides a controlled enhancement of the GV using a spectral distance constraint between the converted and enhanced signals.

3. GLOBAL VARIANCE ENHANCEMENT USING AN LSD CONSTRAINT

The proposed GV enhancement is designed independently of any specific conversion scheme and is applied as a post-processing block. The GV is maximized under a spectral distance constraint, so that the mean distance between the converted vectors (by some conversion method) and their enhanced version is restricted by a pre-set threshold value. This threshold enables the user to control the similarity-variability tradeoff: the GV can be further increased as the similarity is reduced.

Let $\tilde{\mathbf{Y}}_{1:T}$ be a $T \times P$ matrix consisting of a sequence of T converted feature vectors:

$$\tilde{\mathbf{Y}}_{1:T} \triangleq \left(\tilde{\mathbf{y}}_1, \quad \tilde{\mathbf{y}}_2, \quad \ldots \quad , \tilde{\mathbf{y}}_T \right)^\top, \tag{4}$$

where $\{\tilde{\mathbf{y}}_t\}_{t=1}^{T} \in \mathrm{R}^P$ and $(\cdot)^\top$ denotes transpose. We follow [9] and use the Normalized GV (NGV) to measure the dynamic range of the converted spectral features:

$$\mathrm{NGV}\{\tilde{\mathbf{Y}}_{1:T}\} \triangleq \frac{1}{P} \sum_{p=1}^{P} \frac{\mathrm{Var}\{\tilde{\mathbf{Y}}_{1:T}\left(p\right)\}}{\mathrm{Var}\{\mathbf{Y}\left(p\right)\}} \tag{5}$$

where $\mathrm{Var}\{\mathbf{Y}\left(p\right)\}$ is the empirical GV of the p-th element of the target training vectors $\mathbf{Y}\left(p\right) = \{y^q\left(p\right)\}_{q=1}^{Q}$:

$$\mathrm{Var}\{\mathbf{Y}\left(p\right)\} = \frac{1}{Q} \sum_{q=1}^{Q} \left(y^q\left(p\right) - \frac{1}{Q} \sum_{r=1}^{Q} y^r\left(p\right) \right)^2. \tag{6}$$

Let $\tilde{\mathbf{Z}}_{1:T}$ be a $T \times P$ matrix comprising the enhanced version of the converted sequence $\tilde{\mathbf{Y}}_{1:T}$. We set the enhanced sequence as the solution of the following problem:

$$\begin{aligned} \tilde{\mathbf{Z}}_{1:T} &= \underset{\mathbf{Z}_{1:T}}{\mathrm{argmax}} \quad \mathrm{NGV}\{\mathbf{Z}_{1:T}\} \\ &\qquad \text{s.t.} \qquad \overline{\mathrm{LSD}}\left(\mathbf{Z}_{1:T}, \tilde{\mathbf{Y}}_{1:T}\right) \leq \theta_{LSD}, \end{aligned} \tag{7}$$

where $\overline{\mathrm{LSD}}\left(\mathbf{Z}_{1:T}, \tilde{\mathbf{Y}}_{1:T}\right)$ is the mean log-spectral distance (expressed in terms of Mel Frequency Cepstral Coefficients (MFCCs) in (12) below) between the enhanced and converted sequences, $\tilde{\mathbf{Y}}_{1:T}$ and $\mathbf{Z}_{1:T}$, correspondingly, and θ_{LSD} is a pre-set threshold value for the mean LSD in dB. If this threshold is set to zero, the constraint must be satisfied as equality and the converted sequence remains unchanged. For any positive value, the NGV of the enhanced sequence is higher than the NGV of the converted sequence $\tilde{\mathbf{Y}}_{1:T}$, while the $\overline{\mathrm{LSD}}$ between these two sequences is not larger than θ_{LSD}. A naive approach to increase the GV would be to just add white noise to the MFCC parameters with a variance determined by the log-spectral threshold θ_{LSD}. As expected, listening shows that it results in noisy converted speech and is not a viable approach.

We now further develop (7) in terms of explicit expressions for NGV and $\overline{\text{LSD}}$. Define $\mathbf{C}$ as a diagonal $P \times P$ matrix, comprising the GV of the target spectral features evaluated by (6):

$$\mathbf{C} \triangleq diag\left(\text{Var}\{\mathbf{Y}\left(1\right)\}, \text{Var}\{\mathbf{Y}\left(2\right)\}, \ldots, \text{Var}\{\mathbf{Y}\left(P\right)\}\right). \tag{8}$$

Like in [9] we define a covariance operator $\mathbf{\Delta}_T$:

$$\mathbf{\Delta}_T \triangleq \frac{1}{\sqrt{T}}\left(\mathbf{I}_{T\times T} - \frac{1}{T}\mathbf{J}_T\right) \in \mathbb{R}^{T\times T}, \tag{9}$$

where $\mathbf{J}_T$ is a $T \times T$ matrix of all ones. Using (5), (8) and (9) we write the NGV of the converted sequence $\tilde{\mathbf{Y}}_{1:T}$ as:

$$\text{NGV}\{\tilde{\mathbf{Y}}_{1:T}\} = \frac{1}{P}\left\|\Delta_T \cdot \tilde{\mathbf{Y}}_{1:T} \cdot \mathbf{C}^{-\frac{1}{2}}\right\|_2^2. \tag{10}$$

MFCCs are commonly used as spectral features. In this case the mean LSD between each converted vector $\tilde{\mathbf{y}}_t$ and its enhanced version $\tilde{\mathbf{z}}_t$ can be approximated using the Euclidian distance between them:

$$\begin{aligned}\text{LSD}\left(\tilde{\mathbf{z}}_t, \tilde{\mathbf{y}}_t\right) &\approx \frac{10\sqrt{2}}{\ln 10}\sqrt{\sum_{p=1}^{P}\left(\tilde{z}_t\left(p\right) - \tilde{y}_t\left(p\right)\right)^2} \\ &= \kappa\|\tilde{\mathbf{z}}_t - \tilde{\mathbf{y}}_t\|_2 \;\; [dB], \end{aligned} \tag{11}$$

where $\tilde{y}_t\left(p\right)$ and $\tilde{z}_t\left(p\right)$ are the p-th element of the t-th time frame of the converted and enhanced sequences, correspondingly, and $\kappa \triangleq 10\sqrt{2}/\ln 10$.

Therefore, the mean LSD between the two sequences is approximated by:

$$\begin{aligned}\overline{\text{LSD}}\left(\tilde{\mathbf{Z}}_{1:T}, \tilde{\mathbf{Y}}_{1:T}\right) &\approx \frac{\kappa}{T}\sum_{t=1}^{T}\|\tilde{\mathbf{z}}_t - \tilde{\mathbf{y}}_t\|_2 \\ &= \frac{\kappa}{T}\left\|\tilde{\mathbf{Z}}_{1:T} - \tilde{\mathbf{Y}}_{1:T}\right\|_{2,1}, \end{aligned} \tag{12}$$

where $\|\cdot\|_{2,1}$ is the mixed $\ell_{2,1}$ norm [10]. Using (10) and (12), we formulate (7) as:

$$\begin{aligned}\tilde{\mathbf{Z}}_{1:T} &= \underset{\mathbf{Z}_{1:T}}{\text{argmax}} \;\; \left\|\mathbf{\Delta}_T\mathbf{Z}_{1:T}\mathbf{C}^{-\frac{1}{2}}\right\|_2^2 \\ &\quad \text{s.t.} \qquad \left\|\mathbf{Z}_{1:T} - \tilde{\mathbf{Y}}_{1:T}\right\|_{2,1} \le \frac{T\theta_{LSD}}{\kappa}. \end{aligned}$$

We solve the problem by minimizing the Lagrangian:

$$\begin{aligned}\mathcal{L}\left(\mathbf{Z}_{1:T}\right) &= -\left\|\mathbf{\Delta}_T\mathbf{Z}_{1:T}\mathbf{C}^{-\frac{1}{2}}\right\|_2^2 + \\ &+ \lambda\left(\left\|\mathbf{Z}_{1:T} - \tilde{\mathbf{Y}}_{1:T}\right\|_{2,1} - \frac{T\theta_{LSD}}{\kappa}\right). \end{aligned} \tag{13}$$

We obtain the enhanced sequence $\tilde{\mathbf{Z}}_{1:T}$ that minimizes this Lagrangian by numerically evaluating the optimal Lagrange parameter λ^*, as detailed in App. A.

During speech synthesis, each converted sequence is replaced by its GV-enhanced version. Consequently, the GV is increased, while the mean LSD between the enhanced and the originally converted sequence is constrained by $\theta_{LSD}[dB]$.

4. EXPERIMENTAL RESULTS

4.1. Experimental Conditions

Two U.S. English male speakers taken from the CMU ARCTIC database [11] were the source and target speakers. We used 50 parallel sentences for training and 50 other parallel sentences for testing, all sampled at 16kHz and phonetically labeled. The Harmonic Plus Noise Model (HNM) [12] was used for analysis and synthesis by the toolkit available at [13]. MFCCs were used as spectral features ($P = 24$), extracted from the harmonic amplitudes [14]. The analysis frames were time aligned and the feature vectors were matched using a DTW algorithm based on the phonetic labeling [6].

The pitch was converted using a simple linear function using the mean and the standard deviation values of the source and target speakers, $\hat{f}_0^{(y),t}=\mu^{(y)}+\left(\sigma^{(y)}/\sigma^{(x)}\right)\left(f_0^{(x),t}-\mu^{(x)}\right)$, where $f_0^{(x),t}$ and $\hat{f}_0^{(y),t}$ are the pitch values of the source and converted signals at the t-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. In this case the mean and standard deviation of the converted pitch contour match the mean and standard deviation of the pitch values of the target speaker. To reduce audible artifacts, all synthesized waveforms were low-pass filtered at a cut-off frequency of 5kHz.

Three conversion schemes were examined: the classical GMM-based conversion [2], the classical GMM-based conversion followed by the proposed GV enhancement scheme, and CGMM [9]. The synthesized outputs were evaluated using both objective and subjective measures.

4.2. Objective Evaluations

We used two objective measures to evaluate the synthesized outputs: mean Log-Spectral Distortion ($\overline{\text{LSD}}$) between the converted and target signals and normalized GV (NGV). MFCCs were used as spectral features, so the mean LSD between the converted and target signals was evaluated using (12), and the NGV of the converted signals was evaluated using (10). The proposed enhancement method was examined using three threshold values: 1dB, 2dB and 4dB. Several working points were also examined for CGMM, by multiplying the target NGV in the constraint term with factors smaller than 1.

As seen in Table 1, the proposed approach increases the NGV of the converted sentences at the expense of their spectral similarity to the target. Allowing a higher distance between the converted and enhanced signals leads to a further increase of the NGV of the enhanced output. In terms of the objective measures we examined, our method was outperformed by CGMM [9]: for the same NGV of 0.3, CGMM (with a factor) achieved a lower LSD than the proposed approach did, and for the same mean LSD of $7.3dB$, CGMM

Table 1. *Objective performance of the classical GMM-based method (LS-GMM) [2] compared to its enhanced version by the proposed approach and compared to CGMM [9].*

Conversion Method	Mean LSD [dB]	Mean Norm. GV
LS-GMM	6.2	0.1
Enhanced, $\theta_{LSD} = 1dB$	6.4	0.2
CGMM with a factor 0.3	6.4	0.3
Enhanced, $\theta_{LSD} = 2dB$	6.7	0.3
Enhanced, $\theta_{LSD} = 4dB$	7.3	0.4
CGMM	7.3	0.9

achieved a higher NGV than the proposed approach did. However, listening tests, presented in the next subsection, showed that the proposed approach was preferable by the majority of listeners in terms of both similarity and quality, when compared to the other examined approaches, including CGMM.

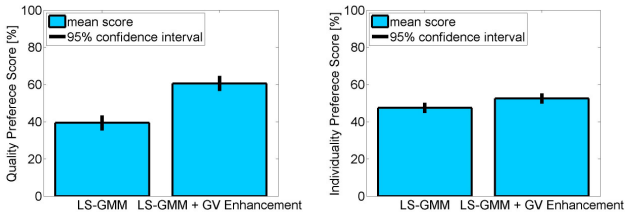
4.3. Subjective Evaluations

Listening tests were carried out to subjectively assess the performance of the examined methods. The examined signals were compared using quality and individuality preference tests. In the quality tests the listeners were asked to indicate the sentence of better quality. In the individuality tests we followed the preceding protocol [8] and the listeners were asked to choose between two possibilities, which of the compared two outcomes is more similar to a given target speaker. In each test, 10 different randomly ordered sentences were examined by 12 listeners (voice samples can be listened to via the link in [15]). The group of listeners comprised 20-30 years old non-experts men and women.

We utilized the controlled enhancement to select the best configuration, in terms of subjective quality. We set $\theta_{LSD} = 2$dB, as informal listening tests showed that the proposed enhancement approach produced the best quality with this threshold value. As mentioned above, several working points were also examined for CGMM using factors smaller than 1 multiplying the target NGV in the constraint term. Eventually, we used the CGMM method with a factor equal to 1 since this value leads to the best quality in our informal listening tests.

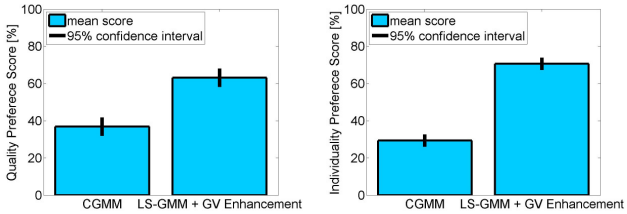
First, we report the impact of the proposed enhancement on the outputs of the classical conversion method [2]. The results, presented in Fig.1(a) show that increasing the GV indeed improved the perceived quality of the converted sentences. Interestingly, the similarity to the target signal was slightly improved, as seen in Fig.1(b), even though objectively, the enhanced signal is less similar to the target speaker in terms of mean LSD.

Second, the overall output of the classical conversion followed by the proposed enhancement was compared to the output of CGMM [9]. The proposed enhancement outperformed CGMM: it was preferred in 60% of the cases in terms of quality and in 70% of the cases in terms of similarity to the target, as seen in Figs.2(a) and 2(b), respectively.



(a) (b)

Fig. 1. The classical GMM conversion method [2] compared with the classical conversion followed by the the proposed enhancement: (a) - quality preference test; (b) - individuality preference test.



(a) (b)

Fig. 2. CGMM [9] compared with the classical GMM conversion followed by the the proposed enhancement: (a) - quality preference test; (b) - individuality preference test.

5. CONCLUSION

One of the main shortcomings of classical spectral voice conversion is that its synthesized outputs sound muffled. Previously proposed approaches deal with the muffling effect by modifying the training process of the conversion.

We propose a new approach for GV enhancement designed independently of any specific conversion method. This method is based on GV maximization under a spectral similarity constraint. The extent of enhancement is controlled by tuning the allowed spectral distance between the enhanced and the originally converted signal. We presented a novel formulation for the mean LSD between two sequences of feature vectors using an $\ell_{2,1}$ norm, so that the threshold value for the spectral distance is specified in [dB].

Experimental results showed that for a given mean LSD, CGMM [9] leads to higher GV than the GV value obtained by the enhancement method proposed here. However, the new enhancement method was selected by the majority of listeners as better than CGMM, both in terms of quality and similarity

to the target. Mean LSD and GV are commonly used for objective evaluation of spectral conversion methods. Still, as shown in this paper, these objective measures do not always agree with subjective evaluations attained by listening tests. Further work is needed to find an alternative measure for objective evaluation of conversion systems, with better correspondence to subjective results.

The proposed enhancement approach was applied here as a post-processing block fed with the outputs of the classical GMM-based conversion. The performance of this new approach could be further examined for other conversion schemes.

A. APPENDIX

We derive the optimal solution that minimizes the Lagrangian in (13), by first diagonalizing the covariance operator, $\mathbf{\Delta}_T = \mathbf{USV}^\top$, and denoting:

$$\mathbf{\Psi} \triangleq \mathbf{V}^\top \mathbf{Z}_{1:T}, \quad \mathbf{\Phi} \triangleq \mathbf{V}^\top \tilde{\mathbf{Y}}_{1:T}, \quad \mathbf{\Omega} \triangleq \mathbf{\Psi} - \mathbf{\Phi} \tag{14}$$

Substituting (14) in (13) we get:

$$\begin{aligned} \mathcal{L}(\mathbf{\Omega}) &= -\left\| \mathbf{S}\left(\mathbf{\Omega} + \mathbf{\Phi}\right) \mathbf{C}^{-\frac{1}{2}} \right\|_2^2 \\ &+ \lambda \left(\|\mathbf{\Omega}\|_{2,1} - \frac{T\theta_{LSD}}{\kappa} \right) \end{aligned} \tag{15}$$

Finally, taking the derivative of (16) with respect to each element of $\mathbf{\Omega}$ and setting it to zero, we obtain the optimal solution:

$$\omega_t(p) = \frac{-\phi_t(p)}{1 - \lambda \mathbf{C}_{p,p} / 2\mathbf{S}_{t,t}^2 \|\boldsymbol{\omega}_t\|_2}. \tag{16}$$

where $\omega_t(p)$ and $\phi_t(p)$ are the (t,p) elements of $\mathbf{\Omega}$ and $\mathbf{\Phi}$, respectively, and $\boldsymbol{\omega}_t = (\omega_t(1), ..., \omega_t(P))^\top$. Since $\|\boldsymbol{\omega}_t\|_2$ depends on $\omega_t(p)$, we use the constraint and set: $\|\boldsymbol{\omega}_t\|_2 = \theta_{LSD}/\kappa$. One of the diagonal elements in the matrix $\mathbf{S}$ is zero, so to avoid ill conditioning we assume, without loss of generality, that it is the last one and evaluate λ using only the first $T-1$ vectors:

$$\sum_{t=1}^{T-1} \|\boldsymbol{\omega}_t\|_2 = \frac{(T-1)\,\theta_{LSD}}{\kappa}. \tag{17}$$

Since $\omega_t(p)$ explicitly depends on the Lagrange parameter λ (see (16)), the optimal value λ^* cannot be simply extracted from (17). Instead, we perform a grid search and set λ^* as the closest value to zero where (17) is approximately sustained. The enhanced sequence is finally obtained by setting $\tilde{\mathbf{Z}}_{1:T}(\lambda^*) = \mathbf{V}\mathbf{\Psi}(\lambda^*)$, where $\mathbf{\Psi}(\lambda^*) = \mathbf{\Omega}(\lambda^*) + \mathbf{\Phi}$.

B. REFERENCES

[1] M. R. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. ICASSP*, 1988, pp. 655–658.

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

[3] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, 1998, pp. 285–288.

[4] A. Kain and M. W. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," in *Proc. ICASSP*, 2001, pp. 813–816.

[5] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of straight spectrum," in *Proc. ICASSP*, 2001, pp. 841–844.

[6] 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.

[7] E. Helander, T. Virtanen, J. Nurminen, and Gabbouj M., "Voice conversion using partial least squares regression," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 18, no. 5, pp. 912921, 2010.

[8] T. Black 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.

[9] H. Benisty and D. Malah, "Voice conversion using GMM with enhanced global variance," in *Proc. Interspeech*, 2011, pp. 669–672.

[10] M. Kowalski and B Torresani, "Sparsity and persistence: mixed norms provide simple signals models with dependent coefficients," *Signal, Image and Video Proc.*, vol. 3, no. 3, pp. 251–264, 2008.

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

[12] 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.

[13] *http://www.talp.cat/talp/index.php/resources/tools/voice-conversion#HSM*.

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

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