Improving the Quality of Standard GMM-Based Voice Conversion Systems by Considering Physically Motivated Linear Transformations

Tudor-Cătălin Zorilă[1,2], Daniel Erro[2], and Inma Hernaez[2]

[1] POLITEHNICA University of Bucharest (UPB), Bucharest, Romania
ztudorc@gmail.com
[2] AHOLAB, University of the Basque Country (UPV/EHU), Bilbao, Spain
{derro,inma}@aholab.ehu.es

Abstract. This paper presents a new method to train traditional voice conversion functions based on Gaussian mixture models, linear transforms and cepstral parameterization. Instead of using statistical criteria, this method calculates a set of linear transforms that represent physically meaningful spectral modifications such as frequency warping and amplitude scaling. Our experiments indicate that the proposed training method leads to significant improvements in the average quality of the converted speech with respect to traditional statistical methods. This is achieved without modifying the input/output parameters or the shape of the conversion function.

Keywords: voice conversion, Gaussian mixture models, dynamic frequency warping, amplitude scaling, linear transformation.

1 Introduction

Voice conversion (VC) has acquired a lot of attention from speech technologies researchers during the last two decades [1–13], being a subject still far from conclusion. VC can be understood as the process by which the voice characteristics of a speaker (source speaker) are replaced by those of another speaker (target speaker) so that the modified speech signal will sound as if it had been produced by the target speaker. VC can be applied to a full range of applications. It can provide an almost costless source of voice variability in text-to-speech (TTS) synthesis, where re-recording new voices is an expensive process and not always possible. This technique can also be applied for voice modifications in movie, music and computer game industries or can be used to repair pathological voices.

VC systems operate in two different modes: training and conversion. During the training phase, given speech recordings from the two involved speakers, the VC systems learn a function to transform the source speaker's acoustic space into that of the target speaker. During the conversion phase, this function is applied to transform new input utterances from the source speaker. Various types of VC techniques have been studied in the literature: vector quantization and mapping codebooks [1], more

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 30–39, 2012.

sophisticated solutions based on fuzzy vector quantization [2], frequency warping transformations [3, 4], artificial neural networks [5], hidden Markov models [6], classification and regression trees [6], etc. However, another technique, namely statistical parametric VC based on Gaussian mixture models (GMM), has prevailed over them.

GMM-based VC systems [7, 8] use statistical principles to partition the acoustic space into a finite number of overlapping classes. Then, a linear transformation is learnt for each class. The function applied during the conversion stage is a statistically weighted combination of these linear transforms. The main problem associated with this well known technique is referred to as oversmoothing. This phenomenon is a consequence of the limited capability of this specific statistical conversion function to capture the correspondence between source and target features in all its variability. As a result of it, the converted speech will sound excesively smoothed and not very natural in terms of subjective quality. Existing methods to alleviate oversmoothing either oversimplify the conversion function [9] or apply sophisticated transformations involving utterance-level features such as the global variance of the converted parameters [10], thus losing the capability of performing frame-by-frame VC in real-time applications.

This paper follows the line of previous works in which frequency warping (FW) based transformations were combined with traditional GMM-based systems [11–13]. FW functions map the frequency axis of the source speaker's spectrum into that of the target speaker. Since they do not remove any detail of the source spectrum, they yield high-quality converted speech judged as quite natural by listeners. However, the conversion accuracy they achieve is moderate because the FW procedure does not modify the relative amplitude of meaningful parts of the spectrum. For this reason, FW was combined with traditional GMM-based systems in several ways [11–13]. In all of these systems, the shape of the VC function had to be modified and more sophisticated signal models and vocoders had to be used to make this combination possible.

In this paper we propose an alternative way of training the set of linear transformations to be applied by a traditional GMM-based VC system. In this new training method, the matrices and vectors of the transformation are calculated according to physical criteria: the matrices are forced to correspond to a FW operation, and the vectors play the role of corrective filters. During conversion, the system operates in the same way as a traditional one and uses the same input/output parameters, i.e. Mel-cepstral coefficients. Despite this, its performance is significantly enhanced in terms of subjective quality, because the degree of oversmoothing is effectively reduced and the converted voice sounds more natural.

The remainder of the paper is structured as follows. Section 2 contains a brief description of the fundamentals of GMM-based voice conversion, including a mathematical interpretation of the oversmoothing effect. In section 3 we show the details of one of the most popular FW training methods. In section 4 we explain the novel training method in which FW-based transformations are integrated into the traditional statistical framework. The effectiveness of this method is experimentally shown in section 5. Finally, the conclusions of this work are summarized in section 6.

2 Traditional GMM-Based VC

The conversion function applied by traditional GMM-based VC systems [7, 8] is a probabilistic combination of m linear transforms:

$$F(\mathbf{x}) = \sum_{i=1}^{m} p_i^{(\theta)}(\mathbf{x}) \left[\mathbf{v}_i + \mathbf{\Gamma}_i \mathbf{\Sigma}_i^{(xx)^{-1}} \left(\mathbf{x} - \boldsymbol{\mu}_i^{(x)} \right) \right] \tag{1}$$

where m is the number of Gaussian mixtures of the model θ, $\boldsymbol{\mu}_i^{(x)}$ and $\mathbf{\Sigma}_i^{(xx)}$ are the mean vector and covariance matrix that characterize the i^{th} Gaussian mixture of θ, and $p_i^{(\theta)}(\mathbf{x})$ is the probability that $\mathbf{x}$ belongs to that specific mixture. Alternatively, the VC function can be expressed as

$$F(\mathbf{x}) = \sum_{i=1}^{m} p_i^{(\theta)}(\mathbf{x}) \left[\mathbf{A}_i \mathbf{x} + \mathbf{b}_i \right] \tag{2}$$

where

$$\mathbf{A}_i = \mathbf{\Gamma}_i \mathbf{\Sigma}_i^{(xx)^{-1}}, \ \mathbf{b}_i = \mathbf{v}_i - \mathbf{A}_i \boldsymbol{\mu}_i^{(x)} \tag{3}$$

Given a training set of paired vectors contained acoustic parameters (Mel-cepstral coefficients in this case), the unknown vectors and matrices of this VC function, $\{\mathbf{v}_i\}$ and $\{\mathbf{\Gamma}_i\}$, can be obtained either by least squares based minimization of the conversion error [7] or by joint density modeling of the concatenated pairs of vectors [8]. In both cases, the resulting converted speech will be perceived by listeners as oversmoothed. Previous investigations on the reasons why oversmoothing appears [9] showed that most of the elements of the matrices $\{\mathbf{A}_i\}$ yielded by traditional training methods were very close to zero due to the limited capability of the GMMs to model the source-target correspondence. In these conditions, the transformation given by expression (1) can be approximated by a simple weighted combination of m vectors $\{\mathbf{v}_i\}$, which explains the observed oversmoothing phenomenon.

In the next section we will show that alternative training methods based on physical principles can provide the traditional linear VC function with matrices and vectors that make it less prone to oversmoothing.

3 Fundamentals of Dynamic Frequency Warping

Dynamic FW (DFW) [3] is a procedure that calculates the FW function that should be applied to a set of (N+1)-point log-amplitude semispectra, $\{X_t\}$, to make them maximally close to their paired counterparts, $\{Y_t\}$. It is based on a cost function $D(i, j)$ which indicates the accumulated log-spectral distortion that would be obtained if the i^{th} bin of the source spectra were mapped into the j^{th} bin of the target spectra following the "best" path from (0, 0) to (i, j). $D(i, j)$ can be expressed mathematically as follows:

$$D(i,j)=\min\begin{Bmatrix}D(i-1,j)+d(i,j)\\ D(i-1,j-1)+w\cdot d(i,j)\\ D(i,j-1)+d(i,j)\end{Bmatrix},\quad i,j=0\ldots N \tag{4}$$

where w, $1 \le w < 2$, is an adjustable weighting coefficient that controls the relative penalty of vertical and horizontal paths ($w \approx 2$ means no penalty for them, while $w \approx 1$ means strong penalty), and $d(i, j)$ is a local distortion measure involving exclusively the i^{th} source bin and the j^{th} target bin. In our implementation, $d(i, j)$ is calculated simultaneously from all the available training vectors to globally optimize the warping procedure:

$$d(i,j)=\sum_{t=1}^{T}\left(X_t[i]-Y_t[j]\right)^2 \tag{5}$$

The frequency warping path P is given by a sequence of points,

$$P=\left\{(0,\ 0),(i_1,j_1),(i_2,j_2),\ldots,(i_K,j_K)\right\}, \tag{6}$$

such that the presence of (i, j) in P indicates that the i^{th} bin of the source spectrum should be mapped into the j^{th} bin of the target spectrum for an optimal warping in terms of log-spectral distortion. In this work, i_K and j_K are forced to be equal to N, so the remaining points of P are backtracked from (N, N) following the minimal-distortion path in inverse order. Note that this path is determined by the recursion in expression (4).

4 Physically Motivated Linear Transforms

DFW is not trainable directly in the parametric domain. Therefore, the first step in the training procedure is translating p^{th}-order cepstral vectors into $(N+1)$-point discrete log-amplitude semispectra. By definition, this can by done by multiplying the cepstral vectors by the following matrix:

$$\mathbf{S}=\begin{bmatrix}1 & 2\cos(\omega_0) & 2\cos(2\omega_0) & \cdots & 2\cos(p\omega_0)\\ 1 & 2\cos(\omega_1) & 2\cos(2\omega_1) & \cdots & 2\cos(p\omega_1)\\ 1 & 2\cos(\omega_2) & 2\cos(2\omega_2) & \cdots & 2\cos(p\omega_2)\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & 2\cos(\omega_N) & 2\cos(2\omega_N) & \cdots & 2\cos(p\omega_N)\end{bmatrix},\quad \omega_k=g\left(k\frac{\pi}{N}\right) \tag{7}$$

where $g(\ )$ is an optional perceptual frequency scale. Note that $g(\ )$ is directly related to the frequency scale assumed during the cepstral analysis.

Similarly, the p^{th}-order cepstral representation of a discrete log-amplitude spectrum can be recovered through the technique known as regularized discrete cepstrum [14],

which is equivalent to multiplying the (N+1)-point discrete log-amplitude semispectrum in vector form by

$$\mathbf{C} = \left(\mathbf{S}^{\mathrm{T}}\mathbf{S} + \lambda\mathbf{R}\right)^{-1}\mathbf{S}^{\mathrm{T}} \tag{8}$$

where $\mathbf{S}$ is given by (7), $\mathbf{R}$ is a regularization matrix that imposes smoothing constraints to the cepstral envelope,

$$R = 8\pi^2 \cdot diag\left\{0, 1^2, 2^2, \ldots p^2\right\}, \tag{9}$$

and λ is an empirical constant typically equal to $2{\cdot}10^{-4}$ [14]. In practice, since the 0^{th} cepstral coefficient (the one carrying the energy) is not considered in voice transformation tasks, we use modified versions of these matrices, $\hat{\mathbf{S}}$ and $\hat{\mathbf{C}}$, where $\hat{\mathbf{S}}$ results from removing the first column of $\mathbf{S}$ and $\hat{\mathbf{C}}$ results from removing the first row of $\mathbf{C}$.

After the training vectors are converted into spectra using matrix $\hat{\mathbf{S}}$, an optimal warping path P is obtained via the DTW training procedure in section 3. Then, we can define the following matrix containing the source-target correspondence:

$$\mathbf{M}[i, j] = \begin{cases} 1, & (i, j) \in P \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

The multiplication of a source semispectrum by $\mathbf{M}^{\mathrm{T}}$ would yield a warped version of the same semispectrum if there were no one-to-many mappings between target and source bins. However, one-to-many mappings are unavoidable according to the structure of P, which is conditioned by the recursion in (4). Therefore, we define the following warping matrix $\mathbf{W}$ in which multiple source bins paired with the same target bin are just averaged:

$$\mathbf{W}[i, j] = \frac{\mathbf{M}[j, i]}{\sum_{k=1}^{N}\mathbf{M}[k, i]} \tag{11}$$

Once $\mathbf{W}$ has been determined, the matrix that converts a p^{th}-order cepstral vector into another cepstral vector representing the warped version of the original spectrum can be easily obtained as

$$\tilde{\mathbf{A}} = \hat{\mathbf{C}} \cdot \mathbf{W} \cdot \hat{\mathbf{S}} \tag{12}$$

Since the frequency response of a corrective filter can be seen as an additive term in the cepstral domain, the cepstral correction vector that is necessary to compensate for the differences between frequency-warped source vectors and target vectors is

$$\tilde{\mathbf{b}} = \mathbf{y}_{avg} - \tilde{\mathbf{A}}\mathbf{x}_{avg} \tag{13}$$

where $\mathbf{x}_{avg}$ and $\mathbf{y}_{avg}$ are computed simply by averaging the source and target cepstral vectors over the training dataset. As a result of this training procedure, we get the following physically motivated linear transformation:

$$\mathbf{y} = \tilde{\mathbf{A}}\mathbf{x} + \tilde{\mathbf{b}} \tag{14}$$

We suggest applying this linear transformation in a traditional statistical framework:

$$F(\mathbf{x}) = \sum_{i=1}^{m} p_i^{(\theta)}(\mathbf{x})\left[\tilde{\mathbf{A}}_i\mathbf{x} + \tilde{\mathbf{b}}_i\right] \tag{15}$$

The matrices and vectors of the transformation can be trained independently for each class of the GMM using exclusively the vectors in that class. For a hard classification, we can assume that $\mathbf{x}$ belongs to the i^{th} class of model θ when $p_i^{(\theta)}(\mathbf{x}) > p_j^{(\theta)}(\mathbf{x})$ for $j = 1\ldots m, j \neq i$. Although such a hard partition of the acoustic space during training is inconsistent with the soft partition used during conversion (15), this does not have any remarkable perceptual consequence according to our listening tests.

5 Experiments and Discussion

The speech data used in the evaluation experiments were taken from the CMU ARCTIC database [15]. Four speakers were selected from this database: two female speakers, *slt* and *clb*, and two male speakers, *bdl* and *rms*. From now on, for the sake of simplicity, they will be referred to as *f*1, *f*2, *m*1 and *m*2, respectively. 50 parallel training sentences per speaker were randomly selected for training and a different set of 50 sentences was separated for testing purposes. The remaining sentences of the database were simply discarded. The sampling frequency of the signals is 16 kHz. We used the vocoder described in [16] to translate the speech signals into Mel-cepstral coefficients and to reconstruct the waveforms from the converted vectors. The order of the cepstral analysis was 24 (plus the 0^{th} coefficient containing the energy, which does not take part in the conversion). The frame shift was set to 8ms. During conversion, the mean and variance of the source speaker's $\log f_0$ distribution were replaced by those of the target speaker by means of a linear transformation. In order to find the correspondence between the source and target cepstral vectors extracted from the parallel training utterances, we calculated a piecewise linear time warping function from the phoneme boundaries given by the available segmentation. The GMMs used in all the experiments had 32 mixtures with full-covariance matrices. Such a number of mixtures was chosen according to phonetic criteria, objective scores measured on separate validation sets, and informal listening tests. During DTW-related computations, N was set to 512.

In the first experiment, different configurations of the proposed method are compared in terms of average Mel-cepstral distortion (MCD) between converted and target vectors. Three specific aspects of the method are studied:

- The influence of the perceptual frequency scale applied when resampling the cepstral envelopes in expression (7). We consider Mel and linear frequency scale. These two configurations will be labeled as "mel" and "lin" respectively.
- The effect of removing the glottal source spectrum from $\{X_t\}$ and $\{Y_t\}$ before training the DFW paths, as suggested in earlier works [3]. In our implementation, we

assume that the glottal spectrum is mainly related with the 1st cepstral coefficient. According to this, we remove the glottal spectrum by settin $c_1 = 0$. This configuration with will be labelled as "c1=0".

— The effect of considering just one representative vector for each class in expression (5), i.e. the average vector, instead of considering all the vectors simultaneously during DFW training. We use labels "avg" and "all" for these configurations.

The MCD scores in Fig. 1, which have been obtained by calculating global scores over all possible combinations of voices, reveal that: (i) considering all the training vectors instead of their average is significantly advantageous; (ii) removing the glottal spectrum is mandatory when only average representative vectors are considered, but it is not crucial when all the vectors are considered during DFW training; (iii) no significant differences can be seen between Mel- and linear-frequency resampling of cepstral envelopes. These observations hold for individual conversion directions.

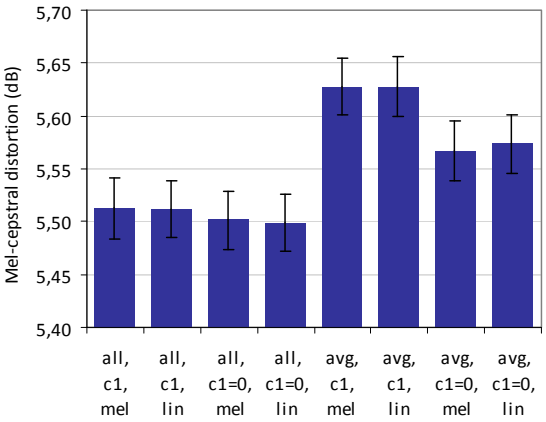

Fig. 1. Average MCD scores and 95% confidence intervals for different configurations of the system and for all combinations of voices

Table 1. Objective comparison between traditional and proposed GMM-based VC systems

	No conversion	Traditional GMM	Proposed GMM
MCD (dB)	7.05 ± 0.03	4.78 ± 0.03	5.50 ± 0.03

Table 1 indicates that a traditional GMM-based system based on joint-density modeling [8] gives significantly better MCD scores than the proposed system regardless of its configuration. Similar observations were made in previous related works [12], where it was also shown that objective distortion measures do not necessary correlate well with subjective measures when the nature of the methods under comparison is heterogeneous. Therefore, we conducted a perceptual mean opinion score (MOS) test to compare the best configuration of the proposed system in terms of

MCD (the one labeled as "all, c1=0, lin") with a traditional GMM-based VC system. In this test, 18 volunteer evaluators listened to reference utterances from the target speakers (previously parameterized and reconstructed with the same vocoder as the converted speech) followed by converted utterances. The listeners were asked to rate the similarity between converted and target voices and the quality of the converted voices in a 5-point scale. As usual, 5 points was the best score and 1 point was the worst. Comparisons were made for 4 different conversion directions: *m*1-*f*1, *f*1-*f*2, *f*2-*m*2, and *m*2-*m*1. The results of the test are shown in Fig. 2. On average, the proposed method significantly outperforms the traditional system in terms of quality while achieving comparable scores in terms of similarity. A more detailed case-by-case analysis reveals that the proposed system is relatively less successful in cross-gender cases. In fact, there is one conversion direction, namely "*f*2-*m*2", in which no quality improvements are achieved. Further analyses indicated that this can be due to the particularities of this specific pair of voices and to some possibly inaccurate decisions regarding the manually adjustable weights and permitted paths in expression (4). These issues will be tackled in future works.

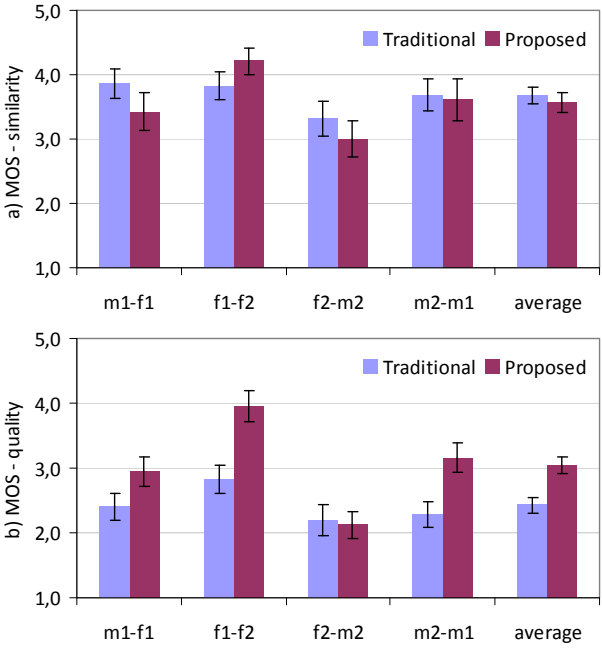

Fig. 2. Mean opinion scores and 95% confidence intervals: a) similarity; b) quality

6 Conclusions

This paper has shown that the performance of traditional voice conversion systems based on Gaussian mixture models and linear transforms can be improved by

imposing some physically meaningful constraints to the matrices and vectors of the transformation. The resulting system is applicable in the same circumstances as the traditional one. Subjective listening tests indicate that on average the proposed method produces evident and statistically significant improvements in quality. Future works will aim at finding the optimal configuration of the system for it to be more robust against the particularities of some specific voice pairs.

Acknowledgements. This work has been partially supported by the Romanian Ministry of Labour, Family and Social Protection (financial agreement POSDRU/88/1.5/S/61178), the Spanish Ministry of Science and Innovation (Buceador, TEC2009-14094-C04-02) and the Basque Government (Berbatek, IE09-262; ZURE_TTS, SPE11UN081).

References

1. Abe, M., Nakamura, S., Shikano, K., Kuwabara, H.: Voice conversion through vector quantization. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 655–658 (1988)
2. Arslan, L.M.: Speaker transformation algorithm using segmental codebooks (STASC). Speech Commun. 28, 211–226 (1999)
3. Valbret, H., Moulines, E., Tubach, J.P.: Voice transformation using PSOLA technique. Speech Commun. 1, 145–148 (1992)
4. Sündermann, D., Ney, H.: VTLN-based voice conversion. In: Proc. IEEE Symp. Signal Process. Inf. Technol., pp. 556–559 (2003)
5. Narendranath, M., Murthy, H.A., Rajendran, S., Yegnanarayana, B.: Transformation of formants for voice conversion using artificial neural networks. Speech Commun. 16(2), 207–216 (1995)
6. Duxans, H., Bonafonte, A., Kain, A., van Santen, J.: Including dynamic and phonetic information in voice conversion systems. In: Proc. Int. Conf. Spoken Lang. Process., pp. 1193–1196 (2004)
7. Stylianou, Y., Cappé, O., Moulines, E.: Continuous probabilistic transform for voice conversion. IEEE Trans. Speech and Audio Process. 6, 131–142 (1998)
8. Kain, A.: High resolution voice transformation. Ph.D. thesis, Oregon Health & Science University (2001)
9. Chen, Y., Chu, M., Chang, E., Liu, J.: Voice conversion with smoothed GMM and MAP adaptation. In: Proc. Eurospeech, pp. 2413–2416 (2003)
10. Toda, T., Black, A.W., Tokuda, K.: Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. IEEE Trans. Audio, Speech, Lang. Process. 15(8), 2222–2235 (2007)
11. Toda, T., Saruwatari, H., Shikano, K.: Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 841–844 (2001)
12. Erro, D., Moreno, A., Bonafonte, A.: Voice conversion based on weighted frequency warping. IEEE Trans. Audio, Speech, Lang. Process. 18(5), 922–931 (2010)
13. Tamura, M., Morita, M., Kagoshima, T., Akamine, M.: One sentence voice adaptation using GMM-based frequency-warping and shift with a sub-band basis spectrum model. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 5124–5127 (2011)

14. Cappé, O., Laroche, J., Moulines, E.: Regularized estimation of cepstrum envelope from discrete frequency points. In: IEEE Workshop on Apps. Signal Process. to Audio & Acoustics, pp. 213–216 (1995)
15. CMU ARCTIC speech synthesis databases, http://festvox.org/cmu_arctic/
16. Erro, D., Sainz, I., Navas, E., Hernaez, I.: HNM-based MFCC+F0 extractor applied to statistical speech synthesis. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 4728–4731 (2011), http://aholab.ehu.es/ahocoder