INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Iterative MMSE Estimation of Vocal Tract Length Normalization Factors for Voice Transformation

Daniel Erro, Eva Navas, Inma Hernáez

AHOLAB, University of the Basque Country (UPV/EHU), Bilbao, Spain
{derro,eva,inma}@aholab.ehu.es

Abstract

We present a method that determines the optimal configuration of a bilinear vocal tract length normalization function to transform the frequency axis of one voice according to a specific target voice. Given a number of parallel utterances of the involved speakers, the single parameter of this function can be calculated through an iterative procedure by minimizing an objective error measure defined in the cepstral domain. This method is also applicable when multiple warping classes are considered, and it can be complemented with amplitude correction filters. The resulting physically motivated cepstral transformation results in highly satisfactory conversion accuracy and improved quality with respect to standard satistical systems.

Index Terms: vocal tract length normalization, voice conversion, frequency warping plus amplitude scaling, speech synthesis.

1. Introduction

State-of-the-art voice conversion (VC) systems are mostly based on statistical methods [1][2], where the conversion functions are defined from mean vectors and covariance matrices of source and target distributions. The inherent limitation of this type of methods is the capability of the GMMs to capture the source-target correspondence for a given parametric representation of speech. The soft classification provided by Gaussian mixture models (GMMs) captures and helps to convert the gross spectral details, while the finer details, which show higher variability inside the acoustic classes, are simply approximated through their mean. This phenomenon is known as oversmoothing. Though modern statistical VC systems provide satisfactory solutions [3], the quality of the converted speech is still far from natural.

Unlike data-driven statistical parametric approaches, frequency warping (FW) functions are based on physical principles. These transformation functions map significant positions of the frequency axis of the source speaker (the central frequencies of formants, for instance) into those of the target speaker [4][5]. As they do not modify the fine spectral details of the source spectrum, they preserve very well the quality of the converted speech. Nevertheless, this is done at the expense of a less accurate conversion when the relative amplitude of different spectral bands is not modified after FW. Consequently, in recent systems [6][7][8] FW is complemented with an amplitude scaling stage to compensate for the conversion inaccuracies.

They way the FW curves are estimated from data is very heterogeneous among existing systems. Non-parametric curves are typically used in the context of VC [4][6][7][8]. In [5][9], however, different types of parametric vocal tract length normalization (VTLN) curves were explored. In these works, the optimal values of the warping parameters were found via grid search. In this paper, we show that the single parameter of a specific type of warping function, namely the bilinear function, can be estimated through an iterative process that minimizes an objective conversion error measure over a set of time-aligned cepstral vectors. Moreover, we show that this method can be extended to a general case in which multiple warping classes are considered. Solutions are provided for both hard and soft classification approaches. Finally, following the line of the aforementioned recent research works on FW, we also propose a method to estimate amplitude scaling filters to increase the accuracy of the FW-based conversion. Subjective experiments reveal that a VC system based on these ideas can give as good performance as a typical GMM-based VC system with similar dimensions in terms of conversion accuracy, while improving the quality of the signals significantly.

The rest of the paper is structured as follows. The new VTLN training method is described in detail in section 2. In section 3 we show how it can be extended to multiple warping classes. Section 4 proposes the complementary amplitude scaling method. The description and results of the subjective evaluation experiments are reported in section 5. Section 6 summarizes the main conclusions of this work.

2. MMSE Estimation of the Warping Factor

All-pass transforms based on bilinear functions are typically applied to perform VTLN [10]. They require just one single parameter α. This transformation is defined in the z domain as

$$z_\alpha^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad |\alpha| < 1 \tag{1}$$

where $z = e^{j\omega}$ and $z_\alpha = e^{j\omega_\alpha}$. The resulting mapping between original and warped frequencies is given by

$$\omega_\alpha = \tan^{-1} \frac{\left(1 - \alpha^2\right)\sin\omega}{\left(1 + \alpha^2\right)\cos\omega - 2\alpha} \tag{2}$$

It is well known that when a cepstral representation of the spectrum is available, it is possible to transform it into a different cepstral sequence that represents the warped spectrum. This cepstral transformation can be expressed as a linear operation [11][12]:

$$\mathbf{y} = \mathbf{W}_\alpha \mathbf{x}, \quad \underset{p\times p}{\mathbf{W}_\alpha} = \begin{bmatrix} 1-\alpha^2 & 2\alpha - 2\alpha^3 & \cdots \\ -\alpha + \alpha^3 & 1 - 4\alpha^2 + 3\alpha^4 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \tag{3}$$

where **x** and **y** are p-dimensional cepstral column vectors. Note that $\mathbf{W}_\alpha$ has been expressed according to the fact that the 0th cepstral coefficient, which carries energy information, is usually excluded from the VC process. As it can be seen from (3), the dependence between $\mathbf{W}_\alpha$ and α is strongly nonlinear. However, α is typically close to zero in VTLN. Therefore, by neglecting the terms of the form α^n for $n>1$ as proposed in [13], this dependence can be linearized:

$$\hat{\mathbf{W}}_\alpha = \begin{bmatrix} 1 & 2\alpha & 0 & 0 & \cdots \\ -\alpha & 1 & 3\alpha & 0 & \cdots \\ 0 & -2\alpha & 1 & 4\alpha & \cdots \\ 0 & 0 & -3\alpha & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (4)$$

The resulting operation $\mathbf{y} = \hat{\mathbf{W}}_\alpha \mathbf{x}$ is equivalent to

$$\mathbf{y} = \mathbf{x} + \alpha \cdot \mathbf{d}(\mathbf{x}) \quad (5)$$

for $\mathbf{d}(\mathbf{x})$ defined as the vector whose ith element is

$$\mathbf{d}(\mathbf{x})[i] = \begin{cases} 2\cdot\mathbf{x}[2] \ , \ i=1 \\ (i+1)\cdot\mathbf{x}[i+1] - (i-1)\cdot\mathbf{x}[i-1] \ , \ i=2...p-1 \\ -(p-1)\cdot\mathbf{x}[p-1] \ , \ i=p \end{cases} \quad (6)$$

These simplified expressions are not accurate enough to perform VTLN by themselves. Nevertheless, they are very advantageous for the estimation of the optimal α from a training dataset. Given a set of N paired p-dimensional cepstral vectors from the source and target speakers, $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_n\}$, the warping factor of the transformation is learned by minimizing the conversion error (MMSE criterion) given by

$$\varepsilon^{(\alpha)} = \sum_{n=1}^{N} \left\| \mathbf{y}_n - \mathbf{W}_\alpha \mathbf{x}_n \right\|^2 \quad (7)$$

If the warping matrix $\mathbf{W}_\alpha$ is approximated through its simplified version in (4)-(6), the error can be rewritten as

$$\varepsilon^{(\alpha)} \cong \sum_{n=1}^{N} \left\| \mathbf{y}_n - \mathbf{x}_n - \alpha\cdot\mathbf{d}(\mathbf{x}_n) \right\|^2 \quad (8)$$

According to (8), it can be shown that the optimal value of α is given by

$$\alpha = \frac{\sum_{n=1}^{N} \mathbf{d}(\mathbf{x}_n)^{\mathrm{T}} \cdot (\mathbf{y}_n - \mathbf{x}_n)}{\sum_{n=1}^{N} \left\| \mathbf{d}(\mathbf{x}_n) \right\|^2} \quad (9)$$

However, in some practical situations, especially in cross-gender VC, α is not so close to zero and the approximation in (4)-(6) is no longer valid. Therefore, we propose an iterative procedure that yields an increasingly accurate solution:

- Step 1: initialize α as 0.
- Step 2: for the current α, calculate a set of warped vectors $\{\mathbf{z}_n\}$, $\mathbf{z}_n = \mathbf{W}_\alpha \mathbf{x}_n$, where the warping matrix is given by expression (3).
- Step 3: calculate the incremental warping factor $\Delta\alpha$ that is necessary to make the vectors $\{\mathbf{z}_n\}$ closer to the target vectors $\{\mathbf{y}_n\}$. This is done by solving (9) for $\{\mathbf{z}_n\}$ instead of $\{\mathbf{x}_n\}$.
- Step 4: accumulate $\Delta\alpha$ into the current α. This can be done via the following well-known expression [14]:

$$\alpha^{(updated)} = \frac{\alpha + \Delta\alpha}{1 + \alpha\cdot\Delta\alpha} \quad (10)$$

- Step 5: if the last $\Delta\alpha$ was insignificant (in other words, if the last value of $\Delta\alpha$ is very close to zero), exit. Otherwise, go back to step 2.

This iterative procedure reaches convergence in a moderate number of iterations (typically less than 25), even in extreme cases of cross-gender VC and for very low amounts of training data.

3. Multiple Warping Classes

Instead of using the same warping factor for the whole signal, in this section we assume m different acoustic classes, each one being characterized by its own factor. The solution is immediate if the classes do not overlap (hard clustering using vector quantization or phonetic labels, for instance), as the method in the previous section can be applied to each class separately. If a soft clustering approach is followed, we can define an instantaneous warping factor as

$$\alpha(\mathbf{x}) = \sum_{k=1}^{m} p_k(\mathbf{x})\alpha_k \quad (11)$$

where α_k is the warping factor ascribed to the kth class and $p_k(\mathbf{x})$ denotes the probability that **x** belongs to that class (not necessarily provided by a statistical model, but also by fuzzy clustering, phonetic clustering combined with fade-in fade-out weights at the boundaries, etc.).

Considering the **x**-dependent warping factor, the set $\{\alpha_k\}$ that minimizes the simplified error in (8) is the least squares solution of the following equation system:

$$\underbrace{\begin{bmatrix} p_1(\mathbf{x}_1)\mathbf{d}(\mathbf{x}_1) & \cdots & p_m(\mathbf{x}_1)\mathbf{d}(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ p_1(\mathbf{x}_N)\mathbf{d}(\mathbf{x}_N) & \cdots & p_m(\mathbf{x}_N)\mathbf{d}(\mathbf{x}_N) \end{bmatrix}}_{Np\times m} \cdot \underbrace{\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix}}_{m\times 1} = \underbrace{\begin{bmatrix} \mathbf{y}_1 - \mathbf{x}_1 \\ \vdots \\ \mathbf{y}_N - \mathbf{x}_N \end{bmatrix}}_{Np\times 1} \quad (12)$$

Similarly as in the single-class case, an iterative method is proposed to compensate for the inaccuracies produced by the simplified expressions.

- Step 1: initialize α_k as 0 for all k.
- Step 2: for the current $\{\alpha_k\}$, calculate a set of warped vectors $\{\mathbf{z}_n\}$, $\mathbf{z}_n = \mathbf{W}_{\alpha(\mathbf{x}_n)} \mathbf{x}_n$, where the warping matrix is given by the combination of expressions (3) and (11).
- Step 3: calculate the incremental warping factors $\{\Delta\alpha_k\}$ that are necessary to make the vectors $\{\mathbf{z}_n\}$ closer to the target vectors $\{\mathbf{y}_n\}$. This is done by solving (12) for $\{\mathbf{z}_n\}$ instead of $\{\mathbf{x}_n\}$ (except for the probabilities, $p_k(\mathbf{x}_n)$, which are kept constant throughout the iterative process to preserve the relationship between factors and classes).
- Step 4: for each k, accumulate $\Delta\alpha_k$ into the current α_k. using expression (10).
- Step 5: if all the last incremental factors in $\{\Delta\alpha_k\}$ are close to 0, exit. Otherwise, go back to step 2.

4. Amplitude Scaling

As mentioned in section 1, recent implementations of frequency warping based systems include an extra module to correct the relative amplitudes of different spectral bands in order to increase the conversion accuracy. This amplitude scaling module is usually implemented in the form of a corrective filter. In our parametric domain, this filter can be seen as a bias vector that is summed to the cepstral representation of the warped spectrum. Without loss of generality, we propose a method to calculate optimal bias vectors in the soft clustering case:

$$\mathbf{b}(\mathbf{x}) = \sum_{k=1}^{m} p_k(\mathbf{x})\mathbf{b}_k \tag{13}$$

Given the warping factors yielded by the iterative method, $\{\alpha_k\}$, the error to be minimized in this step is the following:

$$\varepsilon^{(b)} = \sum_n \left\| \mathbf{r}_n - \mathbf{b}(\mathbf{x}_n) \right\|^2 \;,\quad \mathbf{r}_n = \mathbf{y}_n - \mathbf{W}_{\alpha(\mathbf{x}_n)}\mathbf{x}_n \tag{14}$$

This means calculating the least squares solution of the following equation system:

$$\underset{N\times m}{\begin{bmatrix} p_1(\mathbf{x}_1) & \cdots & p_m(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ p_1(\mathbf{x}_N) & \cdots & p_m(\mathbf{x}_N) \end{bmatrix}} \cdot \underset{m\times p}{\begin{bmatrix} \mathbf{b}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{b}_m^{\mathrm{T}} \end{bmatrix}} = \underset{N\times p}{\begin{bmatrix} \mathbf{r}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{r}_N^{\mathrm{T}} \end{bmatrix}} \tag{15}$$

Although both the warping factors $\{\alpha_k\}$ and the scaling vectors $\{\mathbf{b}_k\}$ are calculated by MMSE in a parametric domain, the resulting conversion can be applied even on the short-time spectrum of a given input signal without restrictions.

5. Evaluation and Discussion

For the evaluation experiments we selected four speakers from the CMU ARCTIC database [15]: two female speakers, *slt* and *clb*, and two male speakers, *bdl* and *rms*. We defined four conversion directions linked to different gender combinations: *bdl-slt*, *slt-clb*, *clb-rms*, and *rms-bdl*. For clarity, they will be referred to as *m-f*, *f-f*, *f-m* and *m-m*, respectively. Fifty parallel training sentences per speaker were selected for training and a different set of fifty sentences were separated for testing purposes. When necessary, the phonetic segmentation of the signals was used to align parallel utterances at frame level via piecewise time warping functions defined from the phoneme boundaries.

To illustrate the performance of the proposed VC method, we compare it with the most widespread statistical method, namely the one presented in [2], which is based on joint density modeling using GMMs. Despite recent advances in GMM-based VC [3], the choice of this baseline method can be justified from the point of view that the conversion function to be applied (a multiplicative matrix and a bias vector, both containing the weighted contribution of each class) is mathematically similar to the one applied by the proposed method (a warping matrix and a scaling vector, both containing the weighted contribution of each class). The comparison is therefore interesting: for a relatively similar formulation, the baseline method is based on statistics, while the proposed one applies a physically motivated transformation. For a fair comparison, we used the same vocoder [16] to get cepstral vectors from the input signals and to reconstruct speech from the converted vectors, and we used the GMM of the baseline method to perform soft clustering in the proposed method. Full-covariance 32-mixture GMMs were used in both methods for soft classification. Such a number of mixtures was chosen according to phonetic information, objective conversion error measures on separate validation sets, and informal listening tests. Both voiced and unvoiced frames were considered.

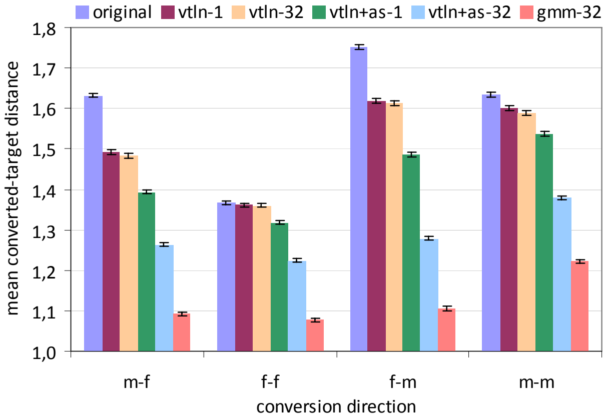

Figure 1: *Mean converted-target distance for different methods and dimensions at 95% confidence intervals.*

Figure 1 shows the mean Euclidean distance between converted and target vectors for several conversion methods: no-modification, single-class and multi-class VTLN with and without amplitude scaling, and GMM-based conversion. We can observe that the contribution of VTLN is remarkable only in cross-gender VC. Apparently, the use of multiple warping classes is not very advantageous with respect to the single-class case. However, informal tests suggest that it leads to local improvements. The contribution of amplitude scaling is clearly visible in all cases. Although the absolute scores achieved by the multi-class VTLN plus amplitude scaling system are worse than the baseline scores, which is not surprising according to previous related works [6][8], subjective tests are necessary to determine whether this is relevant in practice or not.

We carried out a subjective evaluation by means of a mean opinion score (MOS) test in which twenty five volunteers listened to two converted sentences for each method and conversion direction (16 utterances each) and compared them with a reference target utterance (previously parameterized and reconstructed by the vocoder). They rated both the similarity between converted and target voices and the quality of the converted utterance in a 1-to-5 scale. As usual, the best possible score in both performance dimensions was 5.

For each method, Figure 2 shows the MOSs that correspond to the separate conversion directions and also the global average MOS (avg). The performance of the proposed method is comparable to that of the baseline method in terms of average converted-target similarity, and it is much better in terms of average quality. This means that on average the frequency warping plus amplitude scaling procedure is capable of effectively converting voices while preserving the quality of the signals well.

A more detailed case-by-case analysis reveals that the proposed method was relatively less successful in cross-gender

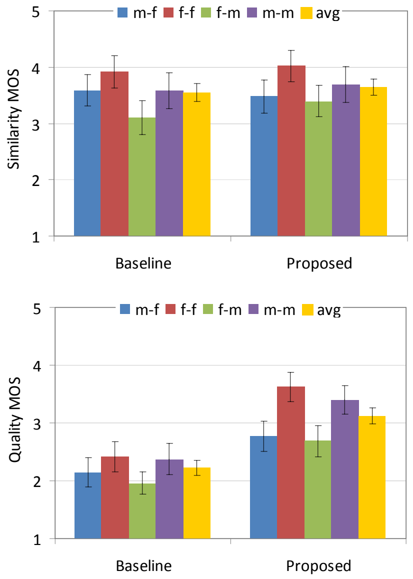

Figure 2: *Results of the subjective test for similarity (top) and quality (bottom). Individual and global mean opinion scores at 95% confidence intervals.*

VC. Indeed, these pair of voices show the highest contrast in vocal tract length. Deeper analyses indicate that the problem is related to the restrictions imposed by the shape of the bilinear functions on the warping capability of the method. During training, whenever the warped source formants and the target formants do not coincide even for the optimal warping factors, the system tries to compensate this mismatch through the bias vectors. Consequently, the resulting vectors contain mixed information not only about the intensity of the formants but also about their location, which may result in less natural transformations. One possible solution might be using more sophisticated warping curves that are applicable on cepstral envelopes, though this would require the MMSE training procedure to be redesigned.

Despite its limitations in cross-gender VC, it is important to observe that the proposed method gives consistently better intra-gender scores than the GMM-based method. Furthermore, even in the cross-gender cases it gives significant quality improvements with respect to the baseline, which uses a relatively similar conversion function with much higher number of parameters. Taking into account that we used the same vocoder in the two systems under evaluation, we can affirm that the quality improvements found in these experiments are not due to the use of sophisticated speech signal models (which partially explains the quality improvements reported in [6], [7] or [8]), but to the mapping method itself.

6. Conclusions

This paper has presented a method to estimate the parameter of a bilinear VTLN function from a set of paired cepstral coefficients by iteratively minimizing the error of the transformation. This method can be applied either for one unique warping class or for multiple classes, even when soft classification is used. A method has been also presented to estimate amplitude scaling terms in accordance with the proposed transformation scheme. Perceptual tests reveal that this parametric frequency-warping plus amplitude-scaling approach yields significant quality improvements with respect to purely statistical methods based on means and covariances, even when they use the same cepstral vocoder.

7. Acknowledgements

This work was supported in part by UPV/EHU (Ayuda de Especialización de Doctores), the Spanish Ministry of Science and Innovation (Buceador Project TEC2009-14094-C04-02), and the Basque Government (Berbatek IE09-262; ZURE_TTS S-PE11UN081).

8. References

[1] Y. Stylianou, O. Cappé, E. Moulines, “Continuous probabilistic transform for voice conversion”, IEEE Trans. Speech and Audio Process., vol. 6, pp. 131-142, 1998.

[2] A. Kain, “High resolution voice transformation”, Ph.D. thesis, Oregon Health & Science University, 2001.

[3] T. Toda, A.W. Black, K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory”, IEEE Trans. Audio, Speech, Lang. Process., vol. 15(8), pp. 2222-2235, 2007.

[4] H. Valbret, E. Moulines, J.P. Tubach, “Voice transformation using PSOLA technique”, Speech Commun., vol. 1, pp. 145-148, 1992.

[5] D. Sündermann, H. Ney, “VTLN-based voice conversion”, Proc. IEEE Symp. Signal Process. Inf. Technol., pp. 556-559, 2003.

[6] D.Erro, A.Moreno, A.Bonafonte, “Voice conversion based on weighted frequency warping”, IEEE Trans. Audio, Speech, Lang. Process., vol. 18(5), pp. 922-931, 2010.

[7] M. Tamura, M. Morita, T. Kagoshima, M. Akamine, “One sentence voice adaptation using GMM-based frequency-warping and shift with a sub-band basis spectrum model”, Proc. ICASSP, pp. 5124-5127, 2011.

[8] E. Godoy, O. Rosec, T. Chonavel, “Spectral envelope transformation using DFW and amplitude scaling for voice conversion with parallel or nonparallel corpora”, Proc. Interspeech, pp. 673-676, 2011.

[9] M. Eichner, M. Wolff, R. Hoffmann, “Voice characteristics conversion for TTS using reverse VTLN”, Proc. ICASSP, pp. 17-20, 2004.

[10] P. Zhan, A. Waibel, “Vocal tract length normalization for large vocabulary continuous speech recognition”, CMU computer science technical reports, 1997.

[11] J. McDonough, W. Byrne, “Speaker adaptation with all-pass transforms”, Proc. ICASSP, pp. 757-760, 1999.

[12] M. Pitz, H. Ney, “Vocal tract normalization equals linear transformation in cepstral space”, IEEE Trans. Speech and Audio Process., vol. 13(5), pp. 930-944, 2005.

[13] T. Emori, K. Shinoda, “Rapid vocal tract length normalization using maximum likelihood estimation”, Proc. Eurospeech, pp. 1649-1652, 2001.

[14] A. Acero, “Acoustical and environmental robustness for automatic speech recognition”, Ph.D. dissertation, Carnegie Mellon Univ., 1990.

[15] Online, “CMU ARCTIC speech synthesis databases”, available in http://festvox.org/cmu_arctic/

[16] D.Erro, I.Sainz, E.Navas, I.Hernaez, “HNM-based MFCC+f0 extractor applied to statistical speech synthesis”, Proc. ICASSP, pp. 4728-4731, 2011. Available at: http://aholab.ehu.es/ahocoder