INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

ML Parameter Generation with a Reformulated MGE Training Criterion – Participation in the Voice Conversion Challenge 2016

D. Erro[1,2], A. Alonso[1], L. Serrano[1], D. Tavarez[1], I. Odriozola[1], X. Sarasola[1], E. Del Blanco[1], J. Sanchez[1], I. Saratxaga[1], E. Navas[1], I. Hernaez[1]

[1]Aholab, University of the Basque Country, Bilbao, Spain
[2]Ikerbasque, Basque Foundation for Science, Bilbao, Spain
derro@aholab.ehu.es

Abstract

This paper describes our entry to the Voice Conversion Challenge 2016. Based on the maximum likelihood parameter generation algorithm, the method is a reformulation of the minimum generation error training criterion. It uses a GMM for soft classification, a Mel-cepstral vocoder for acoustic analysis and an improved dynamic time warping procedure for source-target alignment. To compensate the oversmoothing effect, the generated parameters are filtered through a speaker-independent postfilter implemented as a linear transform in cepstral domain. The process is completed with mean and variance adaptation of the log- fundamental frequency and duration modification by a constant factor. The results of the evaluation show that the proposed system achieves a high conversion accuracy in comparison with other systems, while its naturalness scores are intermediate.
Index Terms: voice conversion, maximum likelihood parameter generation, minimum generation error, linear regression, cepstral postfilter

1. Introduction

A voice conversion (VC) system transforms utterances from a given *source speaker* so as to be perceived as having been uttered by a specific *target speaker*. The VC process consists of two stages (see Fig. 1): (i) training, where the correspondence between source and target acoustic features is learnt from recordings and stored as a conversion function, and (ii) conversion itself, where this function is applied to transform new input utterances from the source speaker. Although the identity of speakers is conveyed not only by segmental features (timbre and average fundamental frequency f_0) but also by suprasegmental (prosody) and even linguistic features, research has been focused mostly on the spectral level.

VC has a relatively long history, along which a wide variety of data-driven conversion function types have been proposed: codebooks [1, 2], hidden Markov models [3, 4, 5], Gaussian mixture models (GMMs) [6, 7, 8, 9, 10, 11], Gaussian processes [12] and shallow/deep neural networks (S/DNNs) [13, 14, 15, 16], among others. Such functions are normally trained from constant-dimension acoustic feature vectors provided by a vocoder. In many other solutions, the conversion function can be applied only to a specific speech signal representation: modification of formant frequencies and bandwidths [17, 18], frequency warping (FW) [19, 20, 21], FW followed by amplitude scaling (AS) [22, 23], etc. When footprint is not an issue, the system can keep some training data from the target speaker and then perform VC through frame selection [24], feature trajectory selection [25], or exemplar-driven transformations [26]. Finally, several hybrid methods have been proposed which exhibit practical advantages in certain situations: fusion of GMM-based VC with FW [27, 28], GMM with unit selection based training [29], FW plus unit selection [30], GMM-driven FW+AS [31, 32], etc.

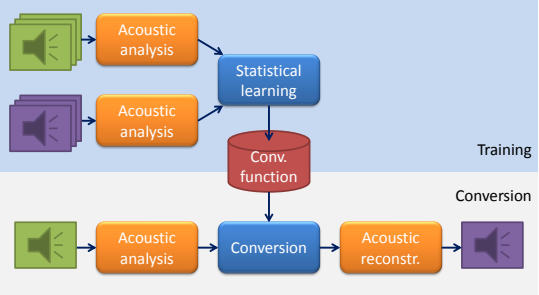

Figure 1: *Block diagram of a statistical VC system. Green: source speaker. Purple: target speaker.*

Given such a wide variety of VC systems/methods and the heterogeneous conditions under which they were originally evaluated, the VC Challenge [33] has been organized to determine how they compare to each other when they are trained with the same data (162 recordings $\times$ 10 different speakers). We considered two different systems for submission. The first one is based on the GMM-driven FW+AS method presented in [32]. The second one is a variant of the maximum likelihood (ML) parameter generation algorithm [9] where the VC parameters, namely a set of linear transforms, are trained through a minimum generation error (MGE) criterion. Our MGE formulation, inspired by classical VC methods [6], differs substantially from that presented in [34]. At the time the Challenge was announced, we were investigating this method in the context of speaker-adaptative HMM-based speech synthesis, but it is easily adaptable to the VC task. Internal subjective evaluations led to the decision of submitting the second system for one main reason: whereas the FW+AS conversion function has relatively few parameters, the alternative one uses unconstrained linear transforms with a greater number of parameters, which seems more adequate given the large amount of training data available. The output of the selected system was enhanced by means of a postfilter and basic prosodic manipulation.

The remainder of this paper is devoted to the description of the method and the analysis of the results of the VC Challenge. Before that, for a better understanding of the theoretical fundamentals of the method, we briefly describe the standard ML parameter generation algorithm in the next section.

http://dx.doi.org/10.21437/Interspeech.2016-219

2. Theoretical background

The well-known ML parameter generation algorithm [9] yields the most probable sequence of p-dimensional acoustic vectors $\{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ given a sequence of Gaussian distributions with mean vectors $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_T\}$ and covariance matrices $\{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_T\}$. These distributions model both the static acoustic parameters and their 1st-order derivative over time. Let us now define $\bar{\mathbf{y}}$ as the supervector that results from concatenating all the (yet unknown) target acoustic vectors:

$$\bar{\mathbf{y}} = [\mathbf{y}_1^\top \ \cdots \ \mathbf{y}_T^\top]^\top \tag{1}$$

We can append derivatives to all the individual vectors in $\bar{\mathbf{y}}$ through the product $\mathbf{W}\bar{\mathbf{y}}$, with $\mathbf{W}$ defined as

$$\mathbf{W} = \mathbf{V} \otimes \mathbf{I} \quad , \quad \mathbf{V} = \begin{bmatrix} 1 & 0 & \cdots & & \\ v[0] & v[1] & \cdots & & \\ 0 & 1 & 0 & \cdots & \\ v[-1] & v[0] & v[1] & \cdots & \\ \cdots & 0 & 1 & 0 & \cdots \\ \cdots & v[-1] & v[0] & v[1] & \cdots \\ & & \vdots & \vdots & \vdots \end{bmatrix} \tag{2}$$

where $\otimes$ denotes the Kronecker product, $\mathbf{I}$ is a $p \times p$ identity matrix, and $\mathbf{V}$ is a $2T \times T$ matrix built from the samples of the window $v[t]$ used to calculate the derivatives[1]. The total log-likelihood of a candidate output supervector $\bar{\mathbf{y}}$ can be expressed as

$$L = -\tfrac{1}{2}(\mathbf{W}\bar{\mathbf{y}} - \bar{\mathbf{u}})^\top \bar{\bar{\mathbf{D}}}(\mathbf{W}\bar{\mathbf{y}} - \bar{\mathbf{u}}) + \kappa \tag{3}$$

where κ is a residual term that does not depend from $\bar{\mathbf{y}}$ and

$$\bar{\mathbf{u}} = [\boldsymbol{\mu}_1^\top \ \ldots \ \boldsymbol{\mu}_T^\top]^\top, \ \ \bar{\bar{\mathbf{D}}} = \text{diag}\left\{\boldsymbol{\Sigma}_1^{-1}, \ldots, \boldsymbol{\Sigma}_T^{-1}\right\} \tag{4}$$

The most likely $\bar{\mathbf{y}}$, i.e. the one that maximizes (3), is

$$\bar{\mathbf{y}} = (\mathbf{W}^\top \bar{\bar{\mathbf{D}}} \mathbf{W})^{-1} \mathbf{W}^\top \bar{\bar{\mathbf{D}}} \bar{\mathbf{u}} \tag{5}$$

Constraining the covariance matrices to be diagonal, the problem can be solved separately for each vector component:

$$\bar{\mathbf{y}}^{(i)} = (\mathbf{V}^\top \bar{\bar{\mathbf{D}}}^{(i)} \mathbf{V})^{-1} \mathbf{V}^\top \bar{\bar{\mathbf{D}}}^{(i)} \bar{\mathbf{u}}^{(i)} \ , \ \ 1 \leq i \leq p \tag{6}$$

where $\bar{\bar{\mathbf{D}}}^{(i)}$ and $\bar{\mathbf{u}}^{(i)}$ contain exclusively the statistics of the i^{th} component and its derivative.

3. Description of the System

The acoustic analysis/reconstruction tool chosen is Ahocoder [35], which parameterizes speech signals by means of a Mel-cepstral (MCEP) representation of the spectral envelope, $\log f_0$ and maximum voiced frequency (MVF). The MVF is related to the harmonicity of the signal and is not modified by our VC system. The 0th MCEP coefficient, related to energy, is neither modified. The next subsections describe how the remaining MCEP coefficients, $\log f_0$ and durations are processed.

3.1. Training

Let us assume N parallel training utterances. First, the source and target MCEP vectors are aligned via dynamic time warping (DTW). Instead of using the standard DTW approach, in order to compensate for the acoustic differences between speakers (mainly in cross-gender conversion), we use an iterative algorithm that can be summarized as follows:

1. Perform classical DTW on every pair of parallel utterances after appending 1st-order derivatives to input vectors.
2. For the current alignment, apply the method in [36] to get the vocal tract length normalization (VTLN) factor α that makes the source vectors closest to the target vectors.
3. If $|\alpha|$ is sufficiently small, take the current alignment as definitive; otherwise, replace the source vectors by their VTLN'ed counterparts ang go to step 1 again.

Similarly as in eq. (1), we group the aligned MCEP vectors into pairs of supervectors $\{\bar{\mathbf{x}}_n, \bar{\mathbf{y}}_n\}_{n=1\ldots N}$, where n is the utterance number. Let us now impose the following relationship between an input source vector $\mathbf{x}$, with time-derivative $\Delta\mathbf{x}$, and the corresponding mean vector $\boldsymbol{\mu}$ used for ML parameter generation:

$$\boldsymbol{\mu} = \sum_{k=1}^{K} \gamma_k(\mathbf{x}) \left(\begin{bmatrix} \mathbf{A}_k & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{A}}_k \end{bmatrix} \cdot \begin{bmatrix} \mathbf{x} \\ \Delta\mathbf{x} \end{bmatrix} + \begin{bmatrix} \mathbf{b}_k \\ \hat{\mathbf{b}}_k \end{bmatrix} \right) \tag{7}$$

where $\gamma_k(\mathbf{x})$ is the probability that $\mathbf{x}$ belongs to the k^{th} class of an acoustic soft-classifier Θ. In this case, Θ is a GMM previously trained from the source training vectors (without derivatives). According to this formulation, the parameters of the VC function can be seen as a matrix

$$\boldsymbol{\Omega} = [\mathbf{A}_1 \ \mathbf{b}_1 \ \hat{\mathbf{A}}_1 \ \hat{\mathbf{b}}_1 \ \ldots \ \mathbf{A}_K \ \mathbf{b}_K \ \hat{\mathbf{A}}_K \ \hat{\mathbf{b}}_K] \tag{8}$$

Assuming that the output of the system will be generated through eq. (6) for every n and that $\bar{\bar{\mathbf{D}}}^{(i)}$ is known (details are given later), we now intend to calculate the optimal $\boldsymbol{\Omega}$ given the pairs $\{\bar{\mathbf{x}}_n, \bar{\mathbf{y}}_n\}_{n=1\ldots N}$. To do this, we suggest using an MGE criterion. The generation error associated to eqs. (6)–(7) is

$$\begin{aligned} \epsilon_i &= \sum_{n=1}^{N} \|\mathbf{Q}_n^{(i)} \boldsymbol{\omega}_i - \bar{\mathbf{y}}_n^{(i)}\|^2 \ , \\ \mathbf{Q}_n^{(i)} &= (\mathbf{V}_n^\top \bar{\bar{\mathbf{D}}}_n^{(i)} \mathbf{V}_n)^{-1} \mathbf{V}_n^\top \bar{\bar{\mathbf{D}}}_n^{(i)} \mathbf{U}(\bar{\mathbf{x}}_n) \end{aligned} \tag{9}$$

where $\boldsymbol{\omega}_i^\top$ is the i^{th} row of $\boldsymbol{\Omega}$ and $\mathbf{U}(\bar{\mathbf{x}})$ (note we omit the utterance number n of supervector $\bar{\mathbf{x}}_n$ for clarity) is equal to

$$\begin{aligned} \mathbf{U}(\bar{\mathbf{x}}) &= \begin{bmatrix} \gamma_1(\mathbf{x}_1)\mathbf{X}_1^\top & \cdots & \gamma_K(\mathbf{x}_1)\mathbf{X}_1^\top \\ \vdots & \ddots & \vdots \\ \gamma_1(\mathbf{x}_T)\mathbf{X}_T^\top & \cdots & \gamma_K(\mathbf{x}_T)\mathbf{X}_T^\top \end{bmatrix} \\ \mathbf{X}_t^\top &= \begin{bmatrix} [\mathbf{x}_t^\top \ 1] & \mathbf{0} \\ \mathbf{0} & [\Delta\mathbf{x}_t^\top \ 1] \end{bmatrix} \end{aligned} \tag{10}$$

The value of $\boldsymbol{\omega}_i$ that minimizes this error can be shown to be

$$\boldsymbol{\omega}_i = \left(\sum_{n=1}^{N} {\mathbf{Q}_n^{(i)}}^\top \mathbf{Q}_n^{(i)} \right)^{-1} \left(\sum_{n=1}^{N} {\mathbf{Q}_n^{(i)}}^\top \bar{\mathbf{y}}_n^{(i)} \right) \tag{11}$$

Thus, the parameters of the VC function $\boldsymbol{\Omega}$ can be calculated on a row-by-row basis. As for $\bar{\bar{\mathbf{D}}}_n^{(i)}$, in this work we use a diagonal matrix built from the variance of component i and its derivative for the GMM class with highest $\gamma_k(\mathbf{x})$ (as Θ does not contain statistics about derivatives, these have to be computed and stored apart). Future works should address the conversion of $\bar{\bar{\mathbf{D}}}_n^{(i)}$ too.

Finally, the mean and variance of $\log f_0$ is computed for both speakers. We also calculate a global duration modification factor D as the inverse of the average DTW slope. To eliminate the influence of initial/final silences, which exhibit irregular durations, instantaneous DTW slopes are weighted by the local energy before averaging.

[1] The derivative of $y[t]$ is $\Delta y[t] = \sum_{\tau=-\infty}^{\infty} v[\tau] y[t+\tau]$. The most typical window is $v[-1] = -\frac{1}{2}, v[1] = \frac{1}{2}, v[t] = 0 \ \forall t \neq \{-1, 1\}$.

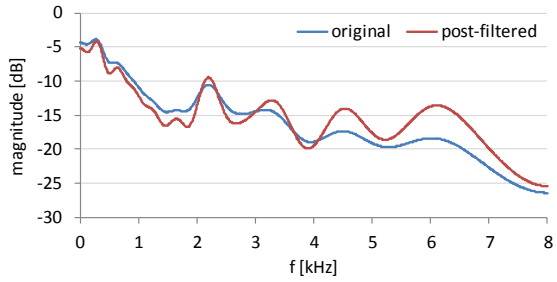

Figure 2: *Effect of postfiltering.*

3.2. Conversion

During conversion, given $\mathbf{\Omega}$ and the GMM Θ, a sequence of mean vectors is derived from the input source MCEP vectors via eq. (7). Then, the corresponding sequence of converted MCEP vectors is generated through eq. (6). The 0^{th} MCEP coefficient is directly copied from the input source vectors. To compensate for the oversmoothing effect, we use a constant (speaker-independent) postfilter. In comparison with other techniques such as global variance enhancement [9], postfiltering has the practical advantage that it is not influenced by the duration of initial/final silences, and also that it can be applied locally at frame level. Moreover, it is compatible with modern parameter generation frameworks based on deep learning [16]. The postfilter applied in this work can be formulated as a linear transform:

$$\mathbf{y}' = \mathbf{P}\mathbf{y} + \mathbf{e} \tag{12}$$

where the multiplicative matrix $\mathbf{P}$ implements a two-band radial postfiltering transform of factors $\{1.03, 1.05\}$ and cutt-off frequency 1 kHz (see [37] for details), and the additive cepstral term $\mathbf{e}$ implements a filter which lifts the mid-high frequencies by 10 dB. The role of $\mathbf{P}$ is sharpening the spectral peaks, and the two-band approach allows for a more intense sharpening at mid-high frequencies, where the oversmoothing effect is more visible. As for $\mathbf{e}$, it enhances the intelligibility of the synthetic speech with no significant quality degradation, according to the findings in [38]. Fig. 2 illustrates the effect of the postfilter. We would like to clarify there was no ad-hoc adjustment of the postfilter parameters (in fact, as discussed later, this configuration was probably not optimal for some speaker pairs).

As for prosody, the mean and variance of $\log f_0$ are rescaled according to the measurements made during training, and the duration of the signal is modified by re-adjusting the frame rate of the vocoder during waveform reconstruction. To avoid unnatural elongation/shortening of the signal, factor D is previously soft-clipped as follows:

$$D' = \exp \frac{\log D}{1 + 2|\log D|} \tag{13}$$

4. Results of the Challenge and Discussion

The proposed system took part in the VC Challenge 2016. The training material provided by the organizers contained 162 parallel recordings from 10 different speakers (5 female + 5 male). Half of them (3 female + 2 male) were selected as source and the remaining ones (2 female + 3 male) were taken as target, which resulted in 25 different conversion pairs with all possible gender combinations. The audio files were released in WAV-mono format at 16 kHz sampling frequency and 16 bits/sample. A few recordings, 12 out of 162, were separated from the training datasets for validation and method selection purposes. During the challenge, 54 new sentences per speaker were released as test dataset.

As mentioned before, the vocoder behind our VC system was Ahocoder [35], and the order of the MCEP parameterization was set to 24. In accordance with phonetic and computational criteria, a 32-component GMM with full covariance matrices was used for training and conversion. Thus, the footprint of a generic VC function was $32 \times (1+24+24 \times 24)$ (from the GMM) + 32×24 (from the diagonal covariances of the derivatives) + $32 \times 2 \times (24 \times 24)$ (from $\mathbf{\Omega}$) + 4 (from $\log f_0$ means and variances) + 1 (from D) = 56869 floating-point numbers, which means less than 500 kB in 'double' precision.

To validate the proposed method (without postfiltering), we compared it with standard GMM-weighted linear regression [6, 8] in terms of MCEP distortion (MCD). For the conversion pairs mentioned above, the proposed method resulted in an average MCD reduction of 2.5%. In fact, MCD reductions between 0.8% and 4.6% were observed for all conversion pairs, with no exception.

The challenge itself consisted of a large-scale perceptual test conducted through a web interface. A total of 200 remunerated evaluators were asked to sit inside a sound-isolated booth, listen to different signals using headphones, and rate two aspects: (i) their naturalness (between 1 = "completely unnatural" and 5 = "completely natural"), and (ii) the similarity to the target speaker ("Same, absolutely sure", "Same, not sure", "Different, not sure", "Different, absolutely sure"). After collecting the individual results of every evaluator, a mean opinion score (MOS) was calculated for naturalness, while the global similarity score was obtained as the percentage of times the speaker was judged to be the same as the target (sure + not sure). The total number of participants was 17 plus a baseline system, namely the VC method in Festvox. For a more detailed explanation of the Challenge, please refer to [33].

The performance of the proposed system is shown in Fig. 3, where scores are compared for each gender combination with the baseline, the mean, the median, the maximum, and the score of the best system in the opposite category. For a more general perspective, in Fig 4 all participants are displayed on a naturalness vs. similarity plane (note the similarity scores have been rescaled). Overall, the proposed system is clearly better than the baseline, mainly in terms of naturalness. There is one case, namely male-to-female conversion, where the baseline is strangely high and outperforms our system, but the baseline's very low quality score reveals that artifacts are hiding its conversion inaccuracies. Except for naturalness in female-to-female conversion, the proposed system is always above the mean. However, given that there are a few systems with very low scores in both performance dimensions (see Fig. 4), the median seems to be a more appropriate reference. The naturalness MOS of the proposed system is exactly the global median score, while the similarity scores are clearly above the median, not far from the best score (according to Fig. 5 only system J achieves a significantly better average similarity score), except for male-to-female conversion. Indeed, we found male-to-female conversion particularly difficult. Given that the other systems' performance does not drop as much as that of our system and that our method has no speaker dependencies, we believe the reason is the use of a constant postfilter: the 10 dB enhancement of the high frequencies was beneficial in terms of intelligibility (not evaluated within this Challenge) but possibly detrimental in terms of naturalness, at least in the male-to-female case. Another possible reason for this performance drop is the harmonic-

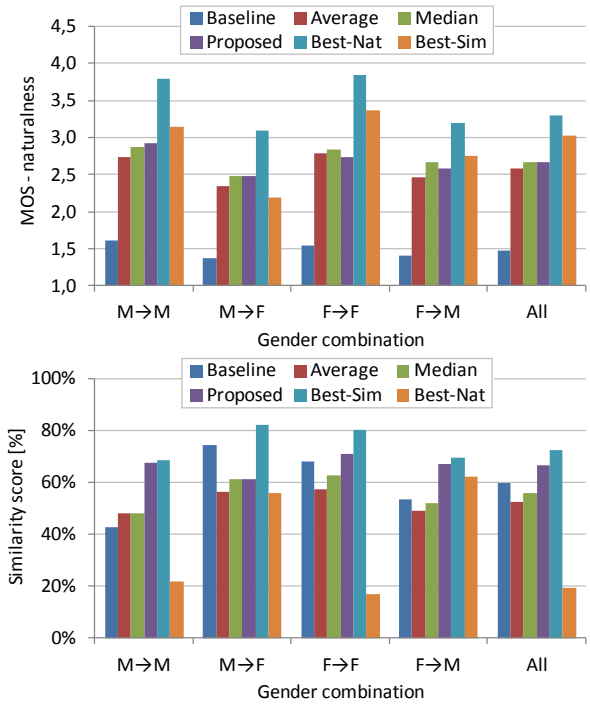

Figure 3: *Analysis of the proposed system's performance. M: male; F: female. Best-Sim: system with the best similarity score. Best-Nat: system with the best naturalness score.*

ity contrast beteen males and females, as we are not converting the MVF. Returning to Fig. 4, if we take the distance r to point (5,5) as global measure of success, five systems (G, J, L, P, O) outperform the proposed one (A) and another one (K) obtains practically the same result ($r = 2.68$).

It was stated in [31] that there is usually a trade-off between the similarity score of a VC system and its naturalness (or quality) score. According to Fig. 4, this trade-off still persists despite the emergence of modern nonlinear mapping methods like DNNs. Interestingly, only one system (J) lies in the rectangle from (3,3) to (4,4). This means that there is still a large room for improvement. Moreover, the system that achieves the largest naturalness score (N) gets also the lowest similarity score. In other words, methods aiming at maximizing the naturalness of the converted speech seem to be paying a high price in terms of conversion accuracy, either because the transform is not flexible enough (as happens with FW) or because higher quality makes conversion inaccuracies more audible. Fortunately, the opposite does not seem to be true, as the winner in terms of similarity (J) ranks 3rd in terms of naturalness.

Regarding the particular method we have used, the results confirm (again) that GMM-based methods, in combination with an adequate post-processing method such as global variance enhancement [9, 11] or postfiltering [37], lead to high similarity scores. We believe we would have obtained slightly better scores if we had converted the MVF and if we had adjusted the postfilter for each conversion pair separately (manual adjustments were not permitted in this Challenge). This encourages us to keep on investigating the use of the suggested MGE training method in speaker-adaptive HMM-based speech synthesis.

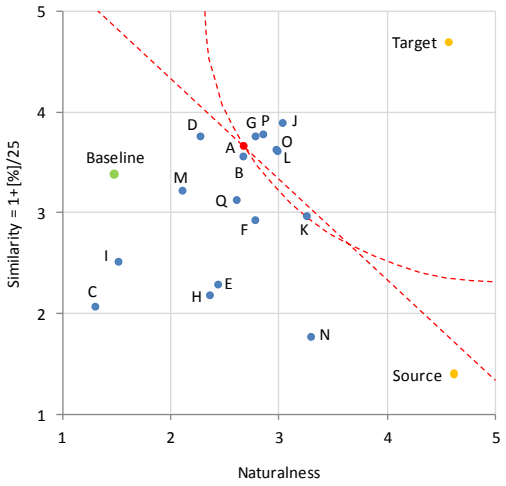

Figure 4: *Results of the VC Challenge in a naturalness vs. similarity plane. The proposed system, A, is marked in red. Straight line: points with the same average score as A. Arc: points with the same distance to (5,5) as A.*

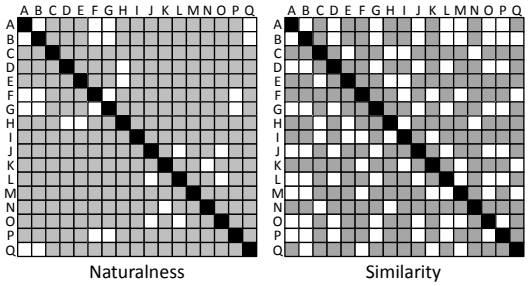

Figure 5: *Statistical significance of the differences between pairs of systems according to pairwise Wilcoxon signed rank tests with Bonferroni correction (level of significance: 1%). Gray: differences are significant; white: they are not.*

5. Conclusions

We have presented a VC system based on ML parameter generation with a reformulated MGE training criterion. This technique can be applied to both VC and speaker-adaptive HMM-based speech synthesis. In combination with mean/variance $\log f_0$ adaptation, constant duration modification and an oversmoothing-compensation postfilter, the method produces quite accurately converted speech with intermediate quality, as shown by the results of the VC challenge 2016.

Future works should focus on MVF conversion and designing a trainable postfilter, as well as extending the MGE method to diagonal covariance matrices.

6. Acknowledgements

This work has been partially funded by the Spanish Ministry of Economy and Competitiveness (RESTORE project, TEC2015-67163-C2-1-R) and the Basque Government (ELKA-ROLA project, KK-2015/00098).

7. References

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

[2] L. M. Arslan, "Speaker transformation algorithm using segmental codebooks (STASC)," *Speech Commun.*, vol. 28, no. 3, pp. 211–226, 1999.

[3] H. Duxans, A. Bonafonte, A. Kain, and J. P. H. van Santen, "Including dynamic and phonetic information in voice conversion systems," in *Proc. Interspeech*, 2004, pp. 1193–1196.

[4] C. H. Lee, C. H. Wu, and J. C. Guo, "Pronunciation variation generation for spontaneous speech synthesis using state-based voice transformation," in *Proc. ICASSP*, 2010, pp. 4826–4829.

[5] H. Zen, Y. Nankaku, and K. Tokuda, "Continuous stochastic feature mapping based on trajectory HMMs," *IEEE Trans. Audio, Speech & Lang. Process.*, vol. 19, no. 2, pp. 417–430, 2011.

[6] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech & Audio Process.*, vol. 6, no. 2, pp. 131–142, 1998.

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

[8] H. Ye and S. J. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," *IEEE Trans. Audio, Speech, & Lang. Process.*, vol. 14, no. 4, pp. 1301–1312, 2006.

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

[10] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Trans. Audio, Speech & Language Process.*, vol. 18, no. 5, pp. 912–921, 2010.

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

[12] N. Xu, Y. Tang, J. Bao, A. Jiang, X. Liu, and Z. Yang, "Voice conversion based on gaussian processes by coherent and asymmetric training with limited training data," *Speech Commun.*, vol. 58, pp. 124–138, 2014.

[13] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," *Speech Commun.*, vol. 16, no. 2, pp. 207–216, 1995.

[14] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 5, pp. 954–964, 2010.

[15] S. H. Mohammadi and A. Kain, "Voice conversion using deep neural networks with speaker-independent pre-training," in *Proc. IEEE SLT*, 2014, pp. 19–23.

[16] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 35–52, 2015.

[17] H. Mizuno and M. Abe, "Voice conversion based on piecewise linear conversion rules of formant frequency and spectrum tilt," in *Proc. ICASSP*, 1994, pp. 469–472.

[18] D. Rentzos, S. Vaseghi, Q. Yan, and C.-H. Ho, "Voice conversion through transformation of spectral and intonation features," in *Proc. ICASSP*, 2004, pp. 21–24.

[19] H. Valbret, E. Moulines, and J. Tubach, "Voice transformation using PSOLA technique," *Speech Commun.*, vol. 11, no. 2-3, pp. 175–187, 1992.

[20] D. Suendermann, G. Strecha, A. Bonafonte, H. Hoege, and H. Ney, "Evaluation of VTLN-based voice conversion for embedded speech synthesis," in *Proc. Interspeech*, 2005, pp. 2593–2596.

[21] Z. Shuang, R. Bakis, S. Shechtman, D. Chazan, and Y. Qin, "Frequency warping based on mapping formant parameters," in *Proc. Interspeech*, 2006, pp. 2290–2293.

[22] M. Tamura, M. Morita, T. Kagoshima, and M. Akamine, "One sentence voice adaptation using gmm-based frequency-warping and shift with a sub-band basis spectrum model," in *Proc. ICASSP*, 2011, pp. 5124–5127.

[23] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Trans. Audio, Speech & Lang. Process.*, vol. 20, no. 4, pp. 1313–1323, 2012.

[24] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Perez, and Y. Stylianou, "Towards a voice conversion system based on frame selection," in *Proc. ICASSP*, 2007, pp. 513–516.

[25] K.-S. Lee, "A unit selection approach for voice transformation," *Speech Commun.*, vol. 60, pp. 30–43, 2014.

[26] Z. Wu, T. Virtanen, E. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Trans. Audio, Speech, & Lang. Process.*, vol. 22, no. 10, pp. 1506–1521, 2014.

[27] 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.

[28] X. Tian, Z. Wu, S. W. Lee, N. Q. Hy, M. Dong, and E. Chng, "System fusion for high-performance voice conversion," in *Proc. Interspeech*, 2015, pp. 2759–2763.

[29] D. Suendermann, H. Hoege, A. Bonafonte, H. Ney, A. W. Black, and S. S. Narayanan, "Text-independent voice conversion based on unit selection," in *Proc. ICASSP*, 2006, pp. 81–84.

[30] Z. Shuang, F. Meng, and Y. Qin, "Voice conversion by combining frequency warping with unit selection," in *Proc. ICASSP*, 2008, pp. 4661–4664.

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

[32] D. Erro, E. Navas, and I. Hernaez, "Parametric voice conversion based on bilinear frequency warping plus amplitude scaling," *IEEE Trans. Audio, Speech, & Lang. Process.*, vol. 21, no. 3, pp. 556–566, 2013.

[33] T. Toda, L. H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The Voice Conversion Challenge 2016," in *Proc. Interspeech*, 2016.

[34] Y. Wu, L. Qin, and K. Tokuda, "An improved minimum generation error based model adaptation for HMM-based speech synthesis," in *Proc. Interspeech*, 2009, pp. 1787–1790.

[35] D. Erro, I. Sainz, E. Navas, and I. Hernáez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal Sel. Topics in Signal Process.*, vol. 8, no. 2, pp. 184–194, 2014.

[36] D. Erro, E. Navas, and I. Hernáez, "Iterative MMSE estimation of vocal tract length normalization factors for voice transformation," in *Proc. Interspeech*, 2012, pp. 86–89.

[37] D. Erro, "Two-band radial postfiltering in cepstral domain with application to speech synthesis," *IEEE Signal Process. Lett.*, vol. 23, no. 2, pp. 202–206, 2016.

[38] M. Koutsogiannaki, P. N. Petkov, and Y. Stylianou, "Intelligibility enhancement of casual speech for reverberant environments inspired by clear speech properties," in *Proc. Interspeech*, 2015, pp. 65–69.