ON COMBINING STATISTICAL METHODS AND FREQUENCY WARPING FOR HIGH-QUALITY VOICE CONVERSION

Daniel Erro, Tatyana Polyakova and Asunción Moreno

TALP Research Center
Universitat Politècnica de Catalunya (UPC)

ABSTRACT

In current voice conversion systems, obtaining a high similarity between converted and target voices requires a high degree of signal manipulation, which implies important quality degradation, up to the point that in some cases the quality scores are unacceptable for real-life applications. Indeed, a tradeoff can be observed between the similarity scores and the quality scores achieved by a given voice conversion system. In our previous works we proved that statistical methods and frequency warping transformations could be combined to yield a better similarity-quality balance than conventional systems, due to significant quality improvements. In this paper, two different ways of combining these two approaches are compared through perceptual tests in order to determine the best strategy for high-quality voice conversion. The comparison is made under the same training conditions, using the same speech model and vector dimensions. The results indicate that the Weighted Frequency Warping method is preferred by listeners.

Index Terms— voice conversion, speech synthesis, gaussian mixture model, weighted frequency warping

1. INTRODUCTION

The goal of voice conversion systems is to modify the voice of a source speaker for it to be perceived as if it had been uttered by another specific speaker, called target speaker. For this purpose, relevant characteristics of the source speaker have to be identified and replaced by those of the target speaker without losing any information or modifying the message. As modifying linguistic features is a very complicated task, most of the existing voice conversion systems focus on the acoustic features of speech. In the area of speech synthesis, voice conversion techniques have important applications. Text-to-speech synthesis (TTS) systems usually generate their output by selecting and concatenating speech units taken from a database, which has been previously built by recording the voice of a professional speaker. Voice conversion technology can be incorporated into TTS systems to transform the recorded voice into any other target voice, so that it would not be necessary to record an entire database for each output voice.

Several voice conversion techniques have been developed since the problem was first formulated in 1988. Abe et al. proposed to convert voices through mapping codebooks created from a parallel training corpus [1]. Arslan tried to avoid the spectral discontinuities caused by the hard partition of the acoustic space by means of a fuzzy classification [2]. Other techniques tried to represent the correspondence between the frequency axis of the source and target speakers by means of a warping function [3]. Due to the low degree of modification, the quality reached by such systems was high, but the conversion scores were not satisfactory because even if the formants were moved to the desired positions, their intensity could not be manipulated. The appearance of statistical methods based on gaussian mixture models (GMM) for spectral envelope transformation was an important breakthrough in voice conversion [4, 5], because the acoustic space of speakers was partitioned into overlapping classes and the weighted contribution of all the classes was considered when transforming acoustic vectors. The spectral envelopes were successfully converted without discontinuities, but in exchange the quality of the converted speech was degraded by over-smoothing. This problem was faced in further works [6, 7, 8], while the usage of GMM-based techniques became almost standard, up to the point that the research was focused on increasing the resolution of GMM-based systems through residual prediction [5, 9, 10] in order to improve both the quality scores and the converted-to-target similarity. Nevertheless, the problem of creating high-quality voice conversion systems that could be used in real-life applications has not been completely solved. At present, there is still a tradeoff between the similarity of converted voices to target voices and the quality achieved by the different conversion methods.

In [11] we presented a new voice conversion technique called Weighted Frequency Warping (WFW), which combined the conversion capabilities of GMM-based systems and the quality of frequency-warping transformations. The aim of WFW was to obtain a better balance between similarity and quality scores than previous existing methods. At the same time, other authors tried to improve conventional GMM-based systems by applying frequency-warping functions to residuals [12]. Both kinds of systems resulted in significant quality improvements and a slight decrement in the converted-to-target similarity scores, although they were conceptually different. This paper compares both approaches by means of a perceptual test, trying to determine the optimal one. For this purpose, both systems were implemented using a common speech model and trained under the same conditions with similar dimensioning parameters, so that the differences observed can be attributed directly to the methods. From now on, our particular implementation of the method combining GMMs and residual frequency warping is called GMM+RWFW for simplicity.

This paper is structured as follows. In section 2, both of voice conversion techniques are explained in detail, emphasizing the differences. In section 3, the results of the subjective test are presented and discussed. Finally, the main conclusions are summarized in section 4.

2. DESCRIPTION OF THE METHODS

First, we observed that WFW and GMM+RWFW can be trained the same way. The training step of both methods consists of estimating a GMM and adequate frequency warping transformations, ideally from a parallel corpus. The differences lie in the way the trained transformation function is applied to the input utterances of the source speaker. In the next subsection, the common training procedure, which is similar to that presented in [11] for WFW, is described. After that, each transformation method is explained in detail.

2.1. Training

Assuming that a parallel (or parallelized) training corpus is available, the acoustic vectors of the source speaker, $\{\mathbf{x}_t\}$, and those of the target speaker, $\{\mathbf{y}_t\}$, are aligned in pairs. Then, a joint-density GMM is estimated from vectors $\{\mathbf{z}_t\}$ by means of the EM algorithm, where $\mathbf{z}_t$ is obtained by concatenating $\mathbf{x}_t$ and $\mathbf{y}_t$. The resulting model is given by the weights $\{\alpha_i\}$, the mean vectors $\{\boldsymbol{\mu}_i\}$ and the covariance matrices $\{\boldsymbol{\Sigma}_i\}$ of its m gaussian components. Individual models for each speaker can be extracted from these parameters, since the mean vectors and covariance matrices can be decomposed into

$$\boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_i^x \\ \boldsymbol{\mu}_i^y \end{bmatrix}, \quad \boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i^{xx} & \boldsymbol{\Sigma}_i^{xy} \\ \boldsymbol{\Sigma}_i^{yx} & \boldsymbol{\Sigma}_i^{yy} \end{bmatrix} \qquad \text{(1a,b)}$$

Once the model is trained, it is possible to calculate the probability that a source vector $\mathbf{x}$ belongs to the i^{th} acoustic class (each gaussian component represents one of the m overlapping acoustic classes):

$$p_i(\mathbf{x}) = \frac{\alpha_i N\left(\mathbf{x}, \boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i^{xx}\right)}{\sum_{j=1}^{m} \alpha_j N\left(\mathbf{x}, \boldsymbol{\mu}_j^x, \boldsymbol{\Sigma}_j^{xx}\right)} \qquad (2)$$

where $N(\cdot)$ denotes a gaussian distribution. In conventional GMM-based methods, each gaussian component is assigned a statistical transformation function, so for a given input vector $\mathbf{x}$ to be converted, the m probabilities $\{p_i(\mathbf{x})\}$ are used as weights for combining the contribution of all the classes:

$$F(\mathbf{x}) = \sum_{i=1}^{m} p_i(\mathbf{x}) \left[\boldsymbol{\mu}_i^y + \boldsymbol{\Sigma}_i^{yx} \boldsymbol{\Sigma}_i^{xx^{-1}} \left(\mathbf{x} - \boldsymbol{\mu}_i^x\right)\right] \qquad (3)$$

More information about GMMs can be found in [4, 5]. On the other hand, in [11] it was proved that high-quality transformations were obtained if optimal frequency warping functions $\{W_i(f)\}$ were calculated for each class. Given an input vector $\mathbf{x}$, the idea was to apply an individual envelope-dependent frequency warping function for converting it, assuming that vectors belonging to the same acoustic class probably required similar warping trajectories:

$$W(\mathbf{x}, f) = \sum_{i=1}^{m} p_i(\mathbf{x}) \cdot W_i(f) \qquad (4)$$

The method proposed for estimating $W_i(f)$ consisted of extracting the formants of the spectral envelopes given by $\boldsymbol{\mu}_i^x$ and $\boldsymbol{\mu}_i^y$, and then searching the correspondence between them in order to establish a piecewise linear frequency warping function. In this paper, the same method has been used to estimate the optimal set of basis frequency warping functions.

2.2. Conversion by WFW

Given a new speech frame to be converted through WFW,

1. The associated vector $\mathbf{x}$ is calculated by parameterizing the spectral envelope.
2. The m weighting probabilities $\{p_i(\mathbf{x})\}$, given by expression (2), are obtained.
3. The individual frequency warping function for the current frame, $W(\mathbf{x}, f)$, is calculated by (4), using the trained basis functions.
4. The magnitude envelope $M(f)$ and the phase envelope $\theta(f)$ of the current frame are warped according to the predicted trajectory $W(\mathbf{x}, f)$:

   $$M'(f) = M\left(W^{-1}(\mathbf{x}, f)\right), \quad \theta'(f) = \theta\left(W^{-1}(\mathbf{x}, f)\right) \qquad \text{(5a,b)}$$

   Note that $\mathbf{x}$ is a low-dimensional parameterized representation of the magnitude envelope, which in general has interesting properties for modeling and transformation, but high-resolution envelopes are required in this step.
5. Finally, the energy distribution of the warped magnitude envelope is corrected by bandwise amplification using the statistically converted envelope $F(\mathbf{x})$ given by expression (3). This is very important for a successful conversion, because it is well known that good

converted-to-target similarity scores cannot be obtained through frequency warping transformations only. The energy correction has to be smooth in order to avoid degrading the quality of the signal. If a very accurate correction was to be applied, the envelopes would be forced to be similar to $F(\mathbf{x})$, so the same performance than typical GMM-based systems would be obtained at the end.

2.3. Conversion by GMM+RWFW

In this case, the spectral envelopes are converted through statistical methods, like in conventional GMM-based systems, but a special treatment is given to residuals. In this context, the word residual denotes the spectral components of the signal that are not captured by the envelope parameterization. Some of these components are due only to codification inaccuracies, and some others are caused by actual high-resolution spectral peaks or valleys that low-order parameterizations are unable to model. This means that moving in frequency this kind of components does not have full physical meaning, but it helps to increase the quality and the perceptual distance between the source speaker and the converted speaker [12]. The way of converting a given input frame is the following:

1. Its associated vector $\mathbf{x}$ is calculated by parameterizing the spectral envelope.
2. The m weighting probabilities $\{p_i(\mathbf{x})\}$, given by expression (2), are obtained.
3. The individual frequency warping function for the current frame, $W(\mathbf{x}, f)$, is calculated by (4).
4. The spectral residual of the current frame is separated by inverse filtering, using the envelope given by $\mathbf{x}$. Then, the warping function obtained in step 3 is applied to the residual:

$$M_r'(f) = M_r\left(W^{-1}(\mathbf{x}, f)\right), \quad \theta_r'(f) = \theta_r\left(W^{-1}(\mathbf{x}, f)\right) \qquad \text{(6a,b)}$$

5. The warped residual is passed through the filter given by the converted envelope $F(\mathbf{x})$, calculated by (3).

Although the algorithm was adapted to the training conditions of WFW, the underlying idea is the same that was proposed in [12].

2.4. Implementation

Both systems were implemented using the same speech model. Due to its flexibility, the harmonic plus stochastic model proposed in [13] was used for analysis, envelope estimation and warping, prosodic modification and reconstruction of speech signals. The harmonic component is represented by f_0, amplitudes and phases. The stochastic component is modelled by means of white noise passing through all-pole filters. The voice conversion algorithms are applied only to the harmonic component of speech, which is present in voiced segments. For this purpose, all-pole filters are fitted to the harmonic amplitudes and are parameterized using line spectral frequencies (LSF), from which GMMs are estimated. The amplitude and phase envelopes to be warped during the conversion step are also extracted from the harmonic parameters, so new harmonic parameters are obtained at the end of the warping process. On the other hand, the unvoiced frames are kept unmodified, whereas the stochastic component of the voiced frames is predicted from the converted harmonic component [11]. Concerning pitch processing, a simple linear transformation based on the log-normal distribution of f_0 has been applied to adapt the pitch range of the source speaker to that of the target speaker:

$$\log f_0' = \mu^y_{\log f_0} + \frac{\sigma^y_{\log f_0}}{\sigma^y_{\log f_0}}\left(\log f_0 - \mu^x_{\log f_0}\right) \qquad (7)$$

This simple approach gives good results when the prosody of the utterances used for testing is neutral and homogeneous. In this case, we used this type of recordings in order to focus the attention of the listeners on the spectral characteristics of the converted voices.

3. EXPERIMENTS

The audio database used for this experiment contained more than 150 sentences in Spanish, uttered by two male and two female speakers. The sampling frequency was 16 KHz and the average duration of the sentences was 4 seconds. 80% of these sentences were used for training the conversion functions. The recorded parallel sentences were aligned for each pair of speakers using HMM-based forced recognition. Concerning the dimensioning of the system, 8th order GMMs were estimated from 14th order LSF vectors. One male and one female speaker were chosen as source, and the other two speakers were used as target, so four different conversion directions were considered: male to male (m2m), female to female (f2f), male to female (m2f) and female to male (f2m). 10 sentences unseen during training were converted and resynthesized for all the combinations of methods and conversion directions, and 28 volunteers were asked to listen to the converted-target sentence pairs in random order. For each pair of voices, listeners were asked to judge if they belonged to the same person using a 5-point scale, from 1 (completely different) to 5 (identical). The final conversion score was obtained by averaging all the individual scores. On the other hand, the listeners were also asked to rate the quality of the converted sentences from 1 point (bad) to 5 points (excellent). The resulting scores are shown in figure 1.

The main differences are found in the quality scores. Although at first sight the naturalness of the utterances converted by GMM+RWFW is not far from that of WFW, the presence of small artifacts introduced by the first method seems to be annoying for the listeners. These artifacts can be due to the interaction between small resonances contained in the residual and the poles of the converted LSF filters. This is probably the main disadvantage of GMM+RWFW: it is very difficult to avoid this kind of harmful interactions because the small spectral peaks of the residuals can be caused simply by codification inaccuracies, so their position is unpredictable.

As we expected before carrying out the test, the conversion scores are slightly better for GMM+RWFW, but the differences found are less significant in this case. It is interesting to observe that, although WFW should in principle achieve worse similarity scores than GMM+RWFW due to the predominance of the frequency warping technique, the results show that in average there is no clear preference.

From a global point of view, as the scores are consistent for all the conversion directions, it can be stated that WFW outperforms GMM+RWFW. Furthermore, the average quality level achieved by WFW is 3.64, which is acceptable for real voice conversion applications.

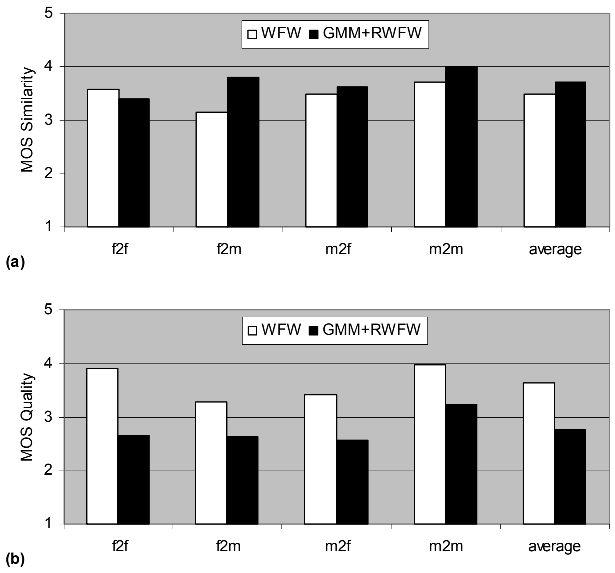

Figure 1: a) Similarity scores. b) Quality scores.

4. CONCLUSIONS

The performance of two different voice conversion techniques combining GMM-based statistical methods and frequency warping transformations has been rated by listeners. The results of the perceptual test indicate that a good balance between similarity and quality scores is obtained by both methods, but significant differences can be observed in the quality scores. The fact that both systems were implemented, trained and dimensioned in the same conditions allows us a more precise evaluation of each method. As the differences should be attributed only to the method itself, we can conclude that the Weighted Frequency Warping technique is more suitable for high-quality voice conversion.

5. ACKNOWLEDGEMENTS

This work was partially funded by the Catalan Government and AVIVAVOZ (TEC2006-13694-C03).

6. REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, H. Kuwabara, "Voice conversion through vector quantization", Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp.655-658, 1988.

[2] L.M. Arslan, "Speaker transformation algorithm using segmental codebooks (STASC)", Speech Communication, no.28, 1999.

[3] H. Valbret, E. Moulines, J.P. Tubach, "Voice transformation using PSOLA technique", Speech Communication, vol.1 pp.145-148, 1992.

[4] Y. Stylianou, O. Cappé, E. Moulines, "Continuous probabilistic transform for voice conversion", Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, vol.6 no.2 pp.131-142, 1998.

[5] A. Kain, "High resolution voice transformation", PhD thesis, OGI School of Science and Engineering, 2001.

[6] T. Toda, H. Saruwatari, K. Shikano, "Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum", Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp.841-844, 2001.

[7] Y. Chen, M. Chu, E. Chang, J. Liu, R. Liu, "Voice conversion with smoothed GMM and MAP adaptation", European Conference on Speech Communications and Technology, pp.2413-2416, 2003.

[8] H. Ye, S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations", IEEE Transactions on Audio, Speech and Language Processing, vol.14, no.4, pp.1301-1312, 2006.

[9] D. Sündermann, H. Höge, A. Bonafonte, H. Duxans, "Residual prediction", Proc. of the IEEE Symposium on Signal Processing and Information Technology, pp.512-516, 2005.

[10] H. Duxans, A. Bonafonte , "Residual conversion versus prediction on voice morphing systems", Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, vol.1, pp.85-88, 2006.

[11] D. Erro, A. Moreno, "Weighted frequency warping for voice conversion", Interspeech'07-Eurospeech, 2007.

[12] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, J. Hirschberg, "TC-Star: cross-language voice conversion revisited", TC-Star Workshop on Speech-to-Speech Translation, 2006.

[13] D. Erro, A. Moreno, A. Bonafonte, "Flexible harmonic/stochastic speech synthesis", 6th ISCA Workshop on Speech Synthesis, 2007.