VOICE CONVERSION USING FRAME SELECTION

Alejandro Uriz, Pablo Daniel Agüero.

Communications Lab
University of Mar del Plata
Argentina

Daniel Erro, Antonio Bonafonte

TALP Research Center
Universitat Politècnica de Catalunya
Spain

ABSTRACT

The goal of voice conversion systems is to modify the voice of a source speaker to be perceived as if it had been uttered by another specific speaker. Many approaches found in the literature introduce an oversmoothing in the target features. Our proposal is the use of features produced by the targer speaker without any smoothing to preserve speaker's identity. The proposed algorithm combines several techniques used in unit selection for text-to-speech. Subjective and objective results support the proposed approach.

Index Terms— speech synthesis

1. INTRODUCTION

The primary goal of voice conversion systems is to modify the voice of a source speaker in order to be perceived as if it had been uttered by another specific speaker: the target speaker. For this purpose, relevant features of the source speaker are identified and replaced by the corresponding features of the target speaker.

In the area of text-to-speech synthesis (TTS) voice conversion techniques play an important role. Since the output voice of TTS is obtained using a large speech database, voice conversion techniques may convert the output into any other target voice by using just a small amount of data to find out the mapping function. The later approach reduces costs and development time.

Several voice conversion techniques have been proposed since the problem was first formulated in 1988. In this year Abe et al. proposed to convert voices through mapping codebooks created from a parallel training corpus [1]. Since then, many authors tried to avoid spectral discontinuities caused by the hard partition of the acoustic space by means of fuzzy classification [2] or frequency axis warping functions [3].

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.

Erro et al. [11] 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.

Speech synthesis with small databases to accomplish voice conversion without a transfer function was studied in Duxans et al. [10]. Although in this case the output speech waveforms were derived directly from the target training data, the identity of the target speaker could not be obtained. The artifacts introduced during the concatenation process (due to the reduced size of the database) degraded the speech signal and made difficult the identification.

Another interesting approach focused in improving target speaker identity is the frame selection approach proposed by Dutoit et al. [13]. In that paper the authors propose to find the optimal sequence of frame target features in training data reducing the distance between source converted features by GMM and target features by means of dynamic programming (the Viterbi algorithm, which was also used in the work of Salor and Demirekler [14]). Sündermann [12] proposed a similar approach just using the source features without any conversion.

In this paper we propose a system that goes back to Abe's proposal, with continuity constraints to avoid concatenation artifacts in speech. The main goal is to maximise the similarity to target speaker by using features extracted from training data, without any smoothing process. The already mentioned over-smoothing of other techniques in the literature produces target features that can not be uttered by the target speaker. In order to compare the system's performance, we made experiments with other state-of-the-art techniques: GMM and WFW.

This paper is organised as follows. In section 2, the three voice conversion techniques are explained in detail, emphasising the differences. In section 3, the results of the subjective test are presented and discussed. Finally, the main conclusions are summarised in section 4.

2. DESCRIPTION OF THE METHODS UNDER STUDY

In this section we describe three methods to perform voice conversion: GMM, WFW, and our proposal, frame selection (FS).

2.1. Voice Conversion using GMM

Assuming that a parallel training corpus is available, the acoustic vectors of the source speaker, x_t, and those of the target speaker, y_t, may be aligned in pairs. Then, a joint-density GMM may be estimated from vectors z_t by means of the EM algorithm, where z_t is obtained by concatenating x_t and y_t. The resulting model is given by the weights p_i, the mean vectors μ_i and the covariance matrices Σ_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

$$\mu_i = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix} \tag{1}$$

$$\mu_i = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix} \tag{2}$$

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

$$p_i(x) = \frac{\alpha N(x, \mu_i^x, \Sigma_i^{xx})}{\sum_{j=1}^{m} \alpha N(x, \mu_j^x, \Sigma_j^{xx})} \tag{3}$$

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 x to be converted, the m probabilities $p_i(x)$ are used as weights for combining the contribution of all the classes:

$$F(x) = \sum_{i=1}^{m} p_i(x) | \mu_i^y + \Sigma_i^{yx} \Sigma_i^{xx\,-1} (x - \mu_i^x) | \tag{4}$$

More information about GMMs can be found in [4, 5], with studies about the dimension of the matrices involved in training. In those papers some simplifications are proposed to reduce the number of parameters and the estimation error, such as diagonal covariance matrices.

2.2. Voice Conversion using WFW

On the other hand, Derro et al. [11] proved that high-quality transformations were obtained if optimal frequency warping functions $W_i(f)$ were calculated for each class. Given an input vector 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:

$$W(x, f) = \sum_{i=1}^{m} p_i(x) W(f) \tag{5}$$

The method proposed for estimating $W_i(f)$ consisted of extracting the formants of the spectral envelopes given by μ_i^x and μ_i^y, and then searching the correspondence between them in order to establish a piecewise linear frequency warping function.

2.3. Voice Conversion using FS

In the literature many methods have been proposed using a transformation function to convert the input source into the target speaker, such as the methods explained in Sections 2.1 and 2.2. Such manipulation of the input vector of features introduces a smoothing, and the converted feature vector may not be realizable by the target speaker. Therefore, the voice conversion produces an unreal feature vector.

In this paper we propose a voice conversion method using frame selection to avoid such effects.

We assume that given a sequence of feature vectors of source speaker (x) we may find an optimal sequence of feature vectors of source speaker in training data ($\hat{x}$) minimising the discontinuities between the corresponding vectors of target speaker in training data ($\hat{y}$):

$$min_{\hat{x}} \left[\sum_i d(x_i, \hat{x}_i) + d(\hat{y}_{i-1}, \hat{y}_i) \right] \tag{6}$$

The formulation assumes that in the first frame ($i = 0$) the concatenation cost $d(\hat{y}_{i-1}, \hat{y}_i)$ is equal to 0.

The method makes the assumption that the sequence of feature vectors of source speaker are correctly aligned with the corresponding feature vectors of target speaker in training data. As a consequence, the voice conversion function is just a lookup table of pairs source-target feature vectors.

In order to take into account the fact that a source feature vector may have many corresponding target feature vectors, we introduce a concatenation cost to minimise discontinuities: $d(\hat{y}_{i-1}, \hat{y}_i)$.

The importance of using a concatenation cost can be explained with an analysis of the dispersion of target feature vectors given a fixed radius of dispersion for source feature vector, as shown in Figure 1. This curve was obtained by randomly selecting a frame of the phone /a/ and searching the k-nearest source feature vectors considering a maximum allowable dispersion. Then, it is possible to calculate the maximum dispersion of the corresponding target feature vectors given a radius of dispersion of source feature vectors.

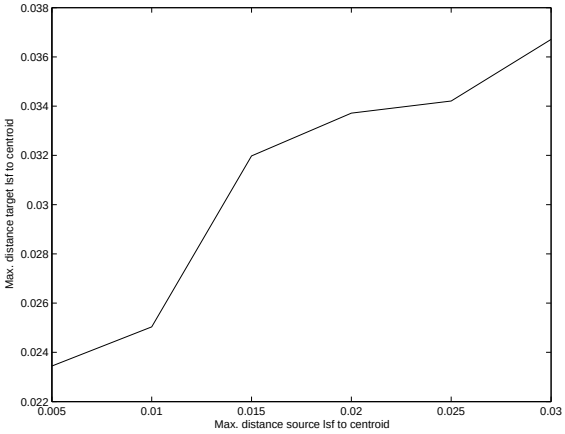

Fig. 1. Results using real contours.

Figure 1 shows that the maximum dispersion for target feature vectors is near 0.023 considering a maximum dispersion for source feature vector of 0.005. Therefore, the dispersion is higher for target feature vectors than for close source feature vectors. This analysis is a measure of the inconsistence of the training data, and the mapping function must take into account that fact. The concatenation cost introduced in our voice conversion method leads the selection of continuous and consistent source feature vectors.

A problem of computational load arises with the proposed conversion method: the size of the search space. The amount of frames in the lookup table is around 120.000 for a 20 minutes database with

a fundamental frequency of 100 Hz (10 ms) using pitch synchronous analysis. As a consequence, to avoid a high number of available frames for each frame i (see Equation 6) of source speaker, we decided to introduce a clustering of source feature vectors to reduce the search space.

Given the source feature vector, the closest centroid with the same phone identity is found. Then, all source feature vectors of the closest centroid and the corresponding target feature vectors are included in the search as candidates. Each phone is divided in three zones: start, medium and end. The phone identity is concatenated with the zone code to preserve the dynamics of phone evolution both for source and target frames.

We can summarise that the training process of the proposed algorithm consists of extracting the parallel feature vectors (14^{th} order LSF vectors) of source and target speakers. Then, a clustering of source feature vectors is performed to reduce the search space in the voice conversion task.

The voice conversion task uses the Viterbi algorithm to obtain an optimal sequence of target feature vectors given the source feature vectors and the target and concatenation costs. Then, the source speaker excitation obtained by inverse filtering is used to synthesise the converted target voice through the converted LPC filters (obtained through an LPC to LSF conversion).

The fundamental frequency contour of the target speaker is obtained with a renormalization in mean and standard deviation of the source speaker contour. Finally, the pitch modification is synthesised using PSOLA.

3. EXPERIMENTS

The audio database used for this experiment contained 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 to train 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, 8^{th} order GMMs were estimated from 14^{th} 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). 35 sentences unseen during training were converted and resynthesized for all methods, and 15 volunteers were asked to listen to the converted-target sentence in random order. Listeners were asked to judge if the voices belonged to the source, target or a third person using a 5-point scale, from 1 (identical to source), 3 (a third person), and 5 (identical to target). 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.

A fourth voice conversion method was included in the experiments. It consists of finding the closest feature vector of target speaker in training data to the real feature vector of target speaker. This voice conversion method that uses privileged information is named **FSopt**. It is a measure of the highest achievable quality and identity by the proposed method.

The results in Figure 2 show the classic trade-off in voice conversion between identity and quality. The methods **WFW** and **GMM** have the highest quality. However, regarding to identity **FSopt** and **FS** have the highest identity, supporting the idea of using real feature vectors instead of smoothed ones.

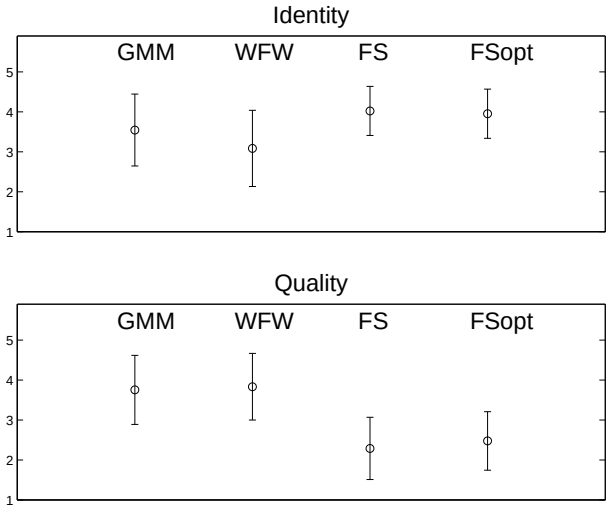

Fig. 2. Identity and quality scores for all systems.

The low quality score of **FSopt** shows that although this method uses privileged information to obtain a high similarity to the LSF parameters of target speaker, the mismatch between vocal tract (target) and excitation (source) reduces the quality.

The small differences between **FS** and **FSopt** show that the dynamic programming algorithm achieve a nearly optimal selection of the sequence of target feature vectors given the source feature vectors.

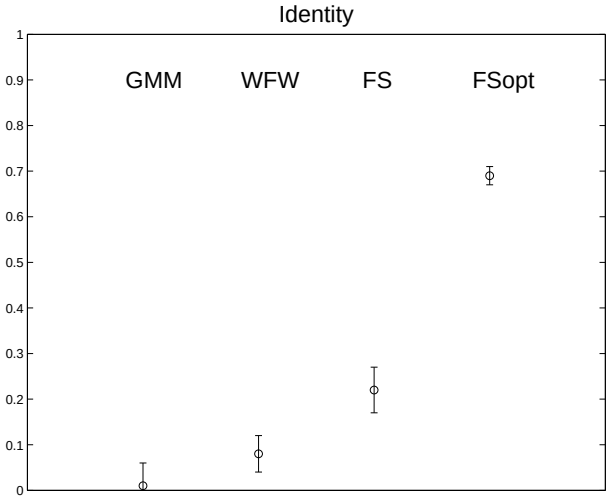

Fig. 3. P distance.

An objective experiment was also conducted to measure the closeness of the converted voice in the four voice conversion methods included in the experiments. We used the P distance included in several works of voice conversion:

$$P = 1 - \frac{d(y, \hat{y})}{d(x, y)} \tag{7}$$

The closer the converted parameters ($\hat{y}$) to the parameters of the target speaker (y), produces that P approaches to one. The distance

between source parameters (x) and target parameters (y) allows to scale the P distance in the virtual path that goes from source to target parameters.

The results in Figure 3 show that **FSopt** is not as close as expected to the target voice, due to missing data in the limited training data. The difference in P parameter between **FSopt** and **FS** shows the margin of improvement available using the proposed technique. However, this difference was not observed into the subjective results, as show in Figure 2. These results support the fact that subjective experiments must always be performed to obtain a real measure of the performance of voice conversion algorithms.

The low scores in **WFW** and **GMM** show identity problems in these techniques introduced by smoothing, as was also shown by the subjective evaluation.

4. CONCLUSIONS

In this paper we presented a voice conversion algorithm that avoids the smoothing effects of other proposals in the literature.

Target and concatenation costs are included in the search of the optimal sequence of target feature vectors given the sequence of source speaker's feature vectors. Objective and subjective results show that the proposed technique achieves high similarity to the target speaker. However, the main drawback is the low quality (2.3) in a five point scale.

Future work will be devoted to improve the quality of the voice conversion introducing a mapping in the excitation. In this paper the excitation extracted from the voice of the source speaker was used to synthesise the voice of the target speaker using target LPC parameters. Mismatches between LPC coefficients and excitation contributed to reduce the final quality.

5. REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1988, pp. 655–658.

[2] L.M. Arslan, "Speaker transformation algorithm using segmental codebooks (STASC)," in *Speech Communication*, 1999, number 28, pp. 655–658.

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

[4] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998, vol. 6, pp. 131–142.

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

[6] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2001, pp. 841–844.

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

[8] H. Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," in *IEEE Transactions on Audio, Speech and Language Processing*, 2006, vol. 14, pp. 1301–1312.

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

[10] H. Duxans and A. Bonafonte, "Residual conversion versus prediction on voice morphing systems," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006, vol. 1, pp. 85–88.

[11] D. Erro and A. Moreno, "Weighted frequency warping for voice conversion," in *Proceedings of Interspeech 2007-Eurospeech*, 2007.

[12] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, "Text-independent voice conversion based on unit selection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006.

[13] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Perez, and Y. Stylianou, "Towards a voice conversion system based on frame selection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[14] Ö.Salor and M. Demirekler, "Voice transformation using principle component analysis based LSF quantization and dynamic programming approach," in *Proceedings of Interspeech 2005*, 2005, pp. 1889–1892.