Frame Alignment Method for Cross-lingual Voice Conversion

Daniel Erro, Asunción Moreno

Department of Signal Theory and Communications
Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
derro@gps.tsc.upc.edu, asuncion@gps.tsc.upc.edu

10.21437/Interspeech.2007-551

Abstract

Most of the existing voice conversion methods calculate the optimal transformation function from a given set of paired acoustic vectors of the source and target speakers. The alignment of the phonetically equivalent source and target frames is problematic when the training corpus available is not parallel, although this is the most realistic situation. The alignment task is even more difficult in cross-lingual applications because the phoneme sets may be different in the involved languages. In this paper, a new iterative alignment method based on acoustic distances is proposed. The method is shown to be suitable for text-independent and cross-lingual voice conversion, and the conversion scores obtained in our evaluation experiments are not far from the performance achieved by using parallel training corpora.

Index Terms: speech synthesis, cross-lingual voice conversion, alignment, GMM, 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 was uttered by another speaker (target speaker). These systems are capable to learn a transformation function from a certain amount of training data of the source and target speakers. In order to map the source speaker's acoustic space to the target speaker's acoustic space, it is necessary to have a previous knowledge about the source-target correspondence between different training units. The process in which this correspondence is established is called alignment, and several strategies for this task have been proposed for different voice conversion systems, depending on the requirements of the training method.

Although some systems based on mapping codebooks [1] or frequency-warping functions [2] try to find a correspondence between acoustic classes of the source and target speakers, most of the voice conversion systems calculate the transformation function from a set of paired parameter vectors, whose correspondence is established at frame level. If a parallel training corpus is available, it is relatively easy to find the correspondence between frames. A parallel corpus contains recordings of the same sentences uttered by both source and target speakers, so that the phonetic content of the paired sentences is the same. In this case, the most preferred frame-alignment technique is dynamic time-warping (DTW), almost standard for parallel corpora [3, 4, 5]. If the orthography and the phonetic transcription are known, speaker-dependent hidden Markov models (HMM) can also be used to segment the utterances. Then, the boundaries of the phonemes or sub-phonemes are taken as anchor points, and time-scale linear interpolation [6] or DTW [7] is used inside the units to establish the correspondence between the source and target vectors.

However, in a realistic voice conversion application, only non-parallel training corpora are available. Four different alignment methods have been used in this situation, and some of them can be used in a cross-lingual context:

- Class mapping [8]. The source and target vectors are classified separately in clusters. A first mapping is established between the acoustic classes of the source and target speakers. Then, the vectors inside each class are mean-normalized and the frame alignment is performed by finding the nearest neighbour of each source vector in the corresponding target class. As the author reports, this method does not provide a high-accuracy alignment.
- Speech recognition [9]. A speech recognizer operating with a speaker-independent HMM is used to label all the source and target frames with a state index. Given the state sequence of one speaker, the alignment procedure consists of finding longest matching sub-sequences from the other speaker until all the frames are paired. The system operates in intra-lingual mode due to the limitation of the recognizer.
- Unit selection using a TTS system [6, 10]. In some applications, a huge database of speech from the source speaker is available, so the TTS system can be employed to generate the same sentences that have been recorded from the target speaker. Thus, a parallel corpus is built from non-parallel recordings. The main disadvantage is the incompatibility with cross-lingual applications and the need of a huge database for synthesis.
- Dynamic programming [11]. Given a set of N source vectors $\{s_k\}$, the dynamic programming technique is used to find the sequence of N target vectors $\{t_k\}$ that minimizes the following cost function:

$$C(\{t_k\}) = \alpha \sum_{k=1}^{N} d(s_k, t_k) + (1-\alpha) \sum_{k=2}^{N} d(t_k, t_{k-1}) \qquad (1)$$

where $d(\,)$ is the acoustic distance between two vectors and α is adjusted depending on the relevance of each term. This alignment technique allows building a text-independent and language-independent system, because the correspondence between frames is obtained using acoustic criteria only. However, it has two drawbacks: (a) it is very time-consuming, and (b) if the training databases are large the conversion scores decay, because the cost function favours the selection of target vectors similar to the source vectors, while some other vectors that contain important specific characteristics of the target speaker do not appear in the sequence $\{t_k\}$ and thus are discarded for the training process.

Some of the systems found in the literature do not actually need an alignment method. Instead, a certain acoustic model is estimated from the data of one of the speakers and the transformation function is calculated using the information provided by the model itself. In [7], acoustic HMMs are trained from the parameter vectors of the target speaker, and a probabilistic voice transformation function is designed in such way that the transformed source vectors give maximum

likelihood with respect to the model. In [12, 13], a conversion function previously trained is fitted to the acoustic data of different speakers using adaptation techniques. Our informal experiments indicate that the conversion functions trained by means of such adaptation methods are less accurate than the reference functions obtained from parallel corpora. On the other hand, adaptation techniques have also been used in HMM-based speech synthesis systems to synthesize speech with different voices [14, 15], but at present the quality of the speech generated by such systems is limited by the synthesis procedure itself.

As the problem of aligning frames for a text-independent cross-lingual voice conversion system has not been completely solved, a new iterative alignment technique based on acoustic distances is proposed in section 2. One of the advantages of the new method is that all the frames of the source and target speakers are considered for the alignment. Thus, all the characteristics of the acoustic space of both speakers are taken into account. The computational load of the method is lower than that of the dynamic programming approach. Our experiments based on perceptual tests indicate that a voice conversion system trained from cross-lingual non-parallel corpora using the proposed aligned method achieves approximately the same results than a similar system trained from intra-lingual parallel corpora. In section 3, a brief description of the voice conversion method used to evaluate the system is given. The new alignment method has been evaluated in a cross-lingual context, as explained in section 4. Finally, the main conclusions are listed in section 5.

2. The New Alignment Method

2.1. Description

The speech recordings of the source and target speakers are analyzed by frames, and the spectral envelope of each frame is parameterized. Let us call $X=\{x_k\}_{k=1...N}$ the set of acoustic parameter vectors of the source speaker, and $Y=\{y_j\}_{j=1...M}$ those of the target speaker. The alignment is carried out as follows:

1. The transformation function is initialized as
$$F(x) = x \qquad (2)$$
2. As a first step, a new set of N transformed vectors X' is created by applying the current transformation function F to the original parameter vectors of X.
$$x'_k = F(x_k), \quad k = 1...N \qquad (3)$$
3. For each vector x'_k in X', the index $p(k)$ of its closest neighbour in Y is found. Similarly, the closest neighbour of each vector y_j is found in X', and the corresponding index is stored as $q(j)$.
4. A new version of the transformation function F is trained from the paired vectors $\{x_k, y_{p(k)}\}$ and $\{x_{q(j)}, y_j\}$. It must be emphasized that the vectors used to train the functions are always those of X and not those of X', which were used only to refine the search of the optimal pairs. As it can be observed, each vector in X is allowed to be paired with more than one vector in Y, and vice versa. Only the repeated pairs are eliminated.
5. Go back to step 2. The process is iterated until convergence is reached.

The transformation function F can be similar to the one used for voice conversion, so that the version of F obtained during the last iteration is directly the solution of the whole conversion problem. In this work, F is chosen to be a GMM-based linear transformation [5] because it gives very good results in terms of objective acoustic distance between converted and target vectors, but any other kind of function that minimizes the error of the transformation between the aligned vectors can be used instead. At each iteration, the parameters $\{\alpha_i, \mu_i, \Sigma_i\}$ of a joint gaussian model of m components are estimated from the concatenated vector pairs $\{[x^T\ y^T]^T\}$. The transformation function is defined by the following equations:

$$F(x) = \sum_{i=1}^{m} p_i(x) \cdot \left[\mu_i^y + \Sigma_i^{yx}\left(\Sigma_i^{xx}\right)^{-1}\left(x - \mu_i^x\right)\right] \qquad (4a)$$

$$p_i(x) = \frac{\alpha_i N\left(x, \mu_i^x, \Sigma_i^{xx}\right)}{\sum_{j=1}^{m} \alpha_j N\left(x, \mu_j^x, \Sigma_j^{xx}\right)} \quad \mu_i = \begin{bmatrix}\mu_i^x \\ \mu_i^y\end{bmatrix} \quad \Sigma_i = \begin{bmatrix}\Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy}\end{bmatrix} \qquad (4b, c, d)$$

where $p_i(x)$ is the probability that a given LSF vector x belongs to the i^{th} gaussian component of the model.

It is desirable that the parameterization used to obtain the vectors of X and Y has good properties for conversion purposes. In typical voice conversion systems, different types of cepstral coefficients or line spectral frequencies (LSF) are usually employed [4, 5]. In this work the alignment has been performed using cepstral coefficients because the cepstral distance between two vectors is a reliable measure of their actual acoustic distance, and this characteristic is important for the 3rd step of the method in which the closest neighbour of each vector is found.

One important advantage of this technique is that all the vectors of the source and target speakers are being used to train the conversion function, so in principle the similarity between converted and target voices is expected to be comparable to that of the parallel-training case.

2.2. Convergence of the method

Although informal perceptual tests show that the method reaches a satisfactory level of alignment, a deeper objective study was carried out in order to prove that the alignment improves at the end of each iteration. For this purpose, a parallel corpus of 100 sentences uttered by four different speakers was used. Two of the speakers were male (m1, m2) and two were female (f1, f2). The average duration of the sentences was approximately 5 seconds. Four different conversion directions were considered in the study: m1-m2, m2-f1, f1-f2 and f2-m1. The corpus was split into two parts: 50 sentences were used for non-parallel alignment and training of the conversion functions (not taking into consideration the fact that they came from a parallel corpus); the rest of the sentences were frame-aligned by HMM used as a reference. Several conversion functions were sequentially trained using an increasing number of alignment iterations. Each of the conversion functions was applied to the reference source vectors, and the cepstral distance was computed between the transformed source vectors and the reference target vectors. Finally, new conversion functions were obtained from the training utterances considering parallel alignments.

Figure 1 shows the results of this experiment. For each of the four conversion directions, the figure plots the accumulated cepstral distance as a function of the number of iterations. As it can be seen, the method is consistent and, in all the conversion directions, the accumulated cepstral distance decreases while the number of iterations increases from 1 to 10. In general, the achieved accumulated distance is not far from the parallel training case, mainly in intra-gender conversion. The case m2-f1 is an exception. Some informal experiments have been carried out to check if other types of initialization for $F(x)$ at the first step lead to better results. In

particular, a frequency warping function has been used for the first iteration. Although no significant improvements have been observed, it is expected that the convergence in cross-gender voice conversion will be reinforced in future works.

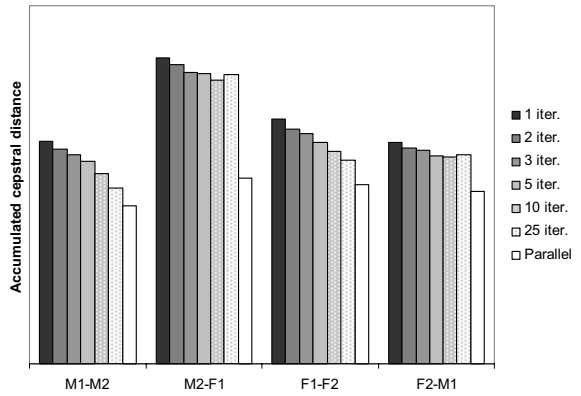

Figure 1: *Objective measurements of accumulated cepstral distance for different number of alignment iterations in intra and cross-gender voice conversion.*

Focusing on the convergence criterion, it can be also observed that there are not very significant improvements from 10 to 25 iterations, while the increment of the computational cost is noticeable. In addition, in cross-gender voice conversion the performance of the system decays when the number of iterations is close to 25. As it is difficult to define an objective convergence criterion, it is reasonable to consider that the performance of the system is accurate enough for 10 alignment iterations. This value has been used for the evaluation of the system, detailed in section 4.

3. The Voice Conversion Method: WFW

As it has been explained in the previous section, GMM-based transformations provide good conversion accuracy between acoustic vectors of different speakers, which is a good property for our alignment method. However, the converted speech has a lower quality because of several factors: over-smoothing and broadening of the formants, effects of the conversion in the analysis/synthesis system (residual, phase spectrum…), etc. Therefore, in real voice conversion applications, listeners may prefer other methods that do not degrade the quality of the converted synthetic speech. In this section a brief description of the WFW voice conversion method used for the evaluation of the alignment system is given. Although the objective acoustic distance between converted and target vectors is higher than that obtained by GMM-based systems, the WFW method is reported to achieve a better balance between the conversion and quality scores when the performance is rated by listeners. More detailed information about this method can be found in [16].

The speech model assumed for the signals is based on the harmonic plus stochastic decomposition, so the frames are represented by the fundamental frequency, the amplitudes and phases of the sinusoids below 5 KHz, and the LPC coefficients of the aperiodic part of the speech frame. Given a set of aligned frame pairs, the spectral envelopes are calculated from the amplitudes of the harmonics by all-pole modeling. A 14^{th} order LSF parameterization is used. An 8^{th} order GMM is trained from the joint source-target LSF vectors and a linear transformation function like that described in (4) is estimated from the model parameters.

Instead of using this linear function to convert the frames, different frequency-warping functions are estimated for each of the acoustic classes given by the GMM. In the conversion phase, the frequency warping trajectory for a given frame is calculated as a linear combination of them. The weights of the linear combination are the probabilities of the current LSF vector to belong to the different acoustic classes (4b), which are provided by the GMM. The amplitudes and phases of the new harmonics are calculated by resampling the warped magnitude and phase envelopes. The converted LSF vector given by the linear GMM transformation (4a) is calculated, and the energy of the warped harmonics is corrected inside some fixed frequency bands using the information provided by the converted LSF envelope. The stochastic component of the voiced frames is predicted from the converted LSF vector using a different linear transformation. No conversion is performed in the unvoiced frames, because it does not lead to significant improvements and it may cause a small loss of quality.

Concerning the prosody, the applied modification consists of linearly transforming the $\log f_0$, using the mean and standard deviation of the log-normal distribution of f_0 extracted from the training data of the source and target speakers.

4. Cross-lingual Evaluation

In the framework of the TC-STAR project, periodical evaluations for all the speech-to-speech translation technologies are organized. In particular, our system participated in the intra-lingual and cross-lingual voice conversion evaluations for Spanish and English. TC-STAR defined an evaluation protocol and the evaluation was carried out by ELDA (Evaluation and Language Resources Distribution Agency). Around 200 sentences in Spanish and 150 in English were recorded from 4 different bilingual speakers, two male speakers (m1, m2) and two female speakers (f1, f2). The average duration of the sentences was 5 seconds. Four different conversion directions were evaluated: m1-m2, m1-f2, f1-m2 and f1-f2. 75% of the sentences were used for the training process and the remaining sentences were analyzed, converted and resynthesized in order to test the performance of the system. The evaluation was based on subjective rating by 20 human judges. All judges were native speakers, and none of them was an expert in speech synthesis. Two metrics were used in the evaluations: one for rating the success of the transformation in achieving the desired speaker identification, and one for rating the quality of the converted speech. To evaluate the conversion degree, the judges were asked to listen to randomly chosen pairs of converted and target sentences and they had to decide using a 5-point scale if the voices came from completely different speakers (1 point) or from exactly the same speaker (5 points). The judges rated the quality of the transformed sentences using a 5-point MOS scale, from bad (1) to excellent (5). Global scores were obtained by averaging the different individual scores. Two systems were tested:

- Intra-lingual with parallel training. Spanish utterances of both, source and target speakers, were used for training and testing. The frame alignment was done with parallel training corpus.
- Cross-lingual Spanish-English. The training data of the source speakers were Spanish sentences, and the training data of the target speakers were in English. The proposed

alignment method was used to train the cross-lingual conversion function.

Table 1 shows the global results of Conversion, Quality and their Average. Looking at the results in Table 1, it can be noticed that the intra-lingual system and the cross-lingual system have similar performances, despite the different training conditions.

Table 2 shows the Conversion scores for each of the conversion directions. Intra-lingual and Cross-lingual behaviors are quite similar for each conversion direction. Cross-gender conversion and specifically the transformation from m1 to f2 is the most problematic in both cases. Notice that the transformation from m1 to m2 is even better in the cross-lingual case than in the intra-lingual with parallel data context. Taking all these facts into account, it can be stated that the new frame alignment method proposed in this paper is suitable for a text-independent cross-lingual voice conversion system.

	Conversion	Quality	Average
Intra-lingual	2.75	2.85	2.80
Cross-lingual	2.63	2.80	2.72

Table 1: *Conversion and Quality scores.*

Conversion	f1-f2	f1-m2	m1-f2	m1-m2
Intra-lingual	2.9	2.9	2.2	3.0
Cross-lingual	2.7	2.3	1.7	3.8

Table 2: *Conversion scores for different conversion directions.*

It must be emphasized that the cross-lingual voice conversion system that used the frame alignment method described in this paper got the best results in the TC-STAR final evaluation among all the evaluated systems.

5. Conclusions

In this paper a new iterative method for frame alignment has been proposed. As the only information required is the acoustic parameterization of the frames, the method can be used to build a text-independent intra- and cross-lingual voice conversion system. Our perceptual tests show that the performance of a voice conversion system using this method to align cross-lingual non-parallel corpora is similar to the one achieved using intra-lingual parallel training corpora.

Future works will focus on the initialization of the method in order to improve the objective and subjective results for cross-gender voice conversion.

6. Acknowledgements

This work was partially supported by TC-STAR (Technology and Corpora for Speech-to-Speech Translation, FP6-506738) and AVIVAVOZ (TEC2006-13694-C03).

7. References

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

[2] D.Sündermann, H.Ney, "VTLN-based voice conversion", Proc. of the IEEE Symposium on Signal Processing and Information Technology, 2003.

[3] 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.

[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] H.Duxans, D.Erro, J.Pérez, F.Diego, A.Bonafonte, A.Moreno, "Voice conversion of non-aligned data using unit selection", TC-STAR Workshop on Speech to Speech Translation, 2006.

[7] 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.

[8] D.Sündermann, A.Bonafonte, H.Ney, H.Höge, "A first step towards text-independent voice conversion", Proc. of the Int. Conf. on Spoken Language Processing, pp.1173-1176, 2004.

[9] H.Ye, S.Young, "Voice conversion for unknown speakers", Proc. of the Int. Conf. on Spoken Language Processing, pp.1161-1164, 2004.

[10] T.En-Najjary, "Conversion de voix pour la synthèse de la parole", PhD thesis, Université de Rennes I, 2005.

[11] D.Sündermann, H.Höge, A.Bonafonte, H.Ney, and J.Hirschberg, "Text-independent cross-language voice conversion", Proc. of the Int. Conf. on Spoken Language Processing, 2006.

[12] A.Mouchtaris, J.Van der Spiegel, P.Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach", IEEE Transactions on Audio, Speech and Language Processing, vol.14 no.3 pp.952-963, 2006.

[13] C.H.Lee, C.H.Wu , "MAP-based adaptation for speech conversion using adaptation data selection and non-parallel training", Proc. of the Int. Conf. on Spoken Language Processing, 2006.

[14] T.Masuko, K.Tokuda, T.Kobayashi, S.Imai, "Voice characteristics conversion for HMM-based speech synthesis system", Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp.1611-1614, 1997.

[15] M.Tamura, T.Masuko, K.Tokuda, T.Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR", Proc. ESCA/COCOSDA Workshop on Speech Synthesis, pp.273-276, 1998.

[16] D.Erro, A.Moreno, "Weighted Frequency Warping for Voice Conversion", Proc. InterSpeech, 2007.