SEGMENT-BASED FRAME ALIGNMENT FOR TEXT-INDEPENDENT VOICE CONVERSION

Zhizheng Wu[1], *Eng Siong Chng*[1], *Haizhou Li*[1,2]

[1]School of Computer Engineering, Nanyang Technological University, Singapore
[2]Human Language Technology Department, Institute for Infocomm Research , Singapore
{wuzz,ASESChng}@ntu.edu.sg, hli@i2r.a-star.edu.sg

ABSTRACT

Most voice conversion techniques require parallel data from the source and target speaker. For the text-independent voice conversion, the most popular approach is that each source frame is paired up with the nearest neighbor target frame in terms of Euclidean distance. In other words, we only consider the spectral similarity between the source and target, ignoring the frame continuity of the resulting target frames. Motivated by similar research in unit-selection of speech synthesis, we proposed a segment-based frame alignment method, which introduces two constraints into the frame pairing between source and target: a) we impose a phonetic constraint that only allows source-target pairing to take place between frames of the same phone segments; b) we take into consideration the target-to-target connectivity when pairing source and target frames. Our objective and subjective experiments on the CMU ARCTIC corpus indicate improvement over conventional frame-based alignment.

Index Terms— voice conversion, text-independent voice conversion, frame alignment

1. INTRODUCTION

The task of voice conversion is to modify one speaker's voice (source) so that it sounds as if it has been uttered by another speaker (target). Voice conversion systems operate on two separate phases, training and conversion phases. In the training phase, the system learns a conversion function between the vocal spaces of the source speaker and target speech from a set of training utterances. The conversion phase takes place at run-time, when the conversion function is applied to the unseen input utterance, and then the converted parameters are passed to a vocoder to reconstruct a speech signal.

A parallel corpus is required to train the conversion function in most of the voice conversion techniques [1, 2, 3], so as to ensure the phonetic content would not be changed during conversion phase. Some conversion algorithms have attempted to relax the requirement of parallel corpus, where the conversion functions are obtained by applying adaptation techniques to existing conversion function[4], however, most of the existing conversion functions are still trained from parallel corpora. It is noted that, unit selection based voice conversion systems have also been proposed in [5, 6], where we select frame in the target speech by minimizing a cost function between a source (or a converted source) and a target frame. These systems avoid frame alignment problem when parallel corpus is not available. Unfortunately they have not addressed adequately the continuity issue of resulting target frames which has been identified in unit selection of speech synthesis research[7].

Many researches have been conducted to the mapping between source and target frames in a non-parallel corpus. The existing simplest alignment method is that each source/target frame is paired up with the nearest neighbor frame in target/source space in terms of Euclidean distance [8], where no phonetic knowledge or contextual information is taken into consideration. This method has been used as the reference baseline[8, 9]. We call it the nearest neighbor frame alignment (NNFA) method in this paper. In [9], a class mapping method is proposed, firstly, the source and target speech vectors are clustered separately into the same number classes. The mapping between the source and the target classes are then established and each source frame vector is paired up with the nearest neighborhood frame vector in corresponding target class in terms of Euclidean distance. Based on the NNFA approach, Erro et al. [8] proposed to train an auxiliary transformation function and generate pseudo parallel database from a non-parallel database, and the auxiliary transformation function and the pseudo parallel database are updated iteratively until convergence.

While Euclidean distance is a good measure between two spectral frames, such frame-level local decision doesn't guarantee the best selection at the utterance level. We believe that the source-target pairing cannot be taken out of phonetic context. A good source-target frame pairing should only happen between source and target segments of the same phonetic labels. We also believe that the target frames selection also needs to ensure a smooth connectivity between the selected frames. In [10], a HMM-based speaker-independent speech recognizer is used to label all the source and target speech frame with a HMM state identity. Give a state sequence of the source speech, the longest matching sequence is found from the target speech. Such a process is repeated until the whole source sequence is paired up with a target sub-sequence. Unfortunately, the acoustic distance between source and target is not exploited in this approach.

In this paper, we continue the quest to find the best frame alignment for text-independent and non-parallel voice conversion. To address the utterance-level continuity issues, we do segment-based frame alignment and further introduce the target-to-target connectivity acoustic distance into the segment pairing cost function. We summarize the process as follows. We first conduct phone-based speech recognition to attain phonetic labels and boundaries for source and target speech. We then select the best matching phonetic segments between source and target according to both source-to-target acoustic distance and target-to-target connectivity. In practice, a Viterbi algorithm is adopted to find the best target segment sequence which minimizes the overall source-to-target and target-to-target acoustic distance.

2. SEGMENT-BASED FRAME ALIGNMENT

In this section, we will briefly introduce the proposed segment-based frame alignment method for text-independent and non-parallel voice conversion.

2.1. Automatic segmentation and transcription

A speaker-independent phoneme recognizer is used to label the source and target speech and identify the phoneme boundaries. When the corpus is small, there could be missing phonemes in either source or target side. In this case, the best solution is to map the phonemes to phonetic classes as shown in Table 1. We will study source-target frame pairing by imposing phonetic class constraint in the experiments.

Table 1. Phoneme to phonetic class mapping [11]

Phonetic classes	Phonemes
Vowel	iy ih eh ey ae aa aw ay ah ao oy ow uh uw er ax ix
Fricative	jh ch s sh z zh f th v dh hh
Nasal	m n ng en
Stop	b d g p t k dx
Approximant	l r w y el
Silence	pause

2.2. Acoustic distance between source and target segments

In this paper, we select the best matching phonetic segments between source and target according to both source-to-target acoustic distance and target-to-target connectivity. The source-to-target acoustic distance is obtained as follows.

First, for each source segment, a short list matching phonetic target segments are pre-selected as candidates, which are the nearest neighbors of the source segment in terms of symmetric Kullback-Leibler (SKL) divergence [12]. See Figure 1: $t_n(1), t_n(2), ..., t_n(K)$ are shortlisted target segment candidates corresponding to the source segment s_n $(1 \leq n \leq N)$.

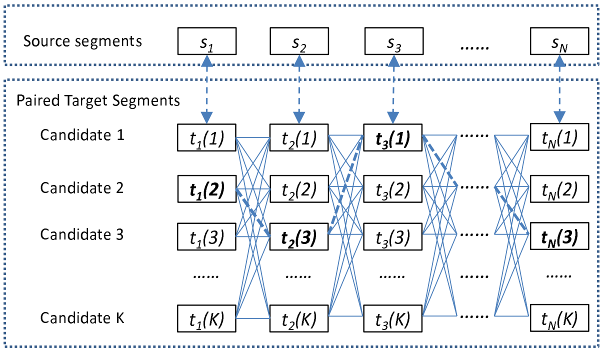

Fig. 1. *Illustration of the proposed segment-based frame alignment approch.*

Then, frame-based dynamic time warping (DTW) is performed between each candidate target segment and the corresponding source segment to identify the frame alignment. As the phoneme recognizer is not perfect, the phoneme boundary may not be exact. In this work, the boundary of source segment is fixed, but both the starting and the ending frame index, l_s and l_e, of the candidate target segment are relaxed. We assume that the starting and ending frame index of a candidate target segment are in the range of $[l_s - 5, l_s + 5]$ and $[l_e - 5, l_e + 5]$, respectively. DTW is performed between the subsequence of one candidate target segment and the corresponding source segment. The subsequence with smallest normalized DTW distance is chosen as a 'real' candidate target segment. This normalized DTW distance, which is the average Euclidean distance of the source-target frame pairs, is used as the source-to-target acoustic distance between a source and target segment.

2.3. Boundary connectivity between target segments

To ensure the target-to-target connectivity, which is ignored in the frame based alignment, we propose a boundary connectivity distance which is defined as follows:

$$\begin{aligned} d(t_{n-1}(i), t_n(j)) \quad = \quad & \frac{1}{3}[d(y_l^{t_{n-1}(i)}, y_1^{t_n(j)}) + d(y_{l-1}^{t_{n-1}(i)}, y_1^{t_n(j)}) \\ & + d(y_l^{t_{n-1}(i)}, y_2^{t_n(j)})] \qquad (1) \end{aligned}$$

where $d(t_{n-1}(i), t_n(j))$ is the boundary connectivity distance between the $t_{n-1}(i)$ and $t_n(j)$ segments. $t_{n-1}(i)$ and $t_n(j)$ are the rank i and j target candidate segments in the short list corresponding to the source segments s_{n-1} and s_n, respectively. (See Fig. 1). $d(y_l^{t_{n-1}(i)}, y_1^{t_n(j)})$ is the Euclidean distance of the last frame in the $t_{n-1}(i)$ segment and the first frame in the $t_n(j)$ segment. One notes that if the $t_{n-1}(i)$ and $t_n(j)$ segments are actually neighbors in the target speech, their boundary connectivity distance will be very small.

2.4. Search of the optimal target segment sequence

Give a source segment sequence $s_1, s_2, ..., s_N$, K target segment candidates are pre-selected in terms of SKL distance measure, and the source-to-target , target-to-target boundary connectivity distance are also obtained by segmental DTW and Eq. (1), respectively. Then the optimal target segment sequence $t_1(q_1), t_2(q_2), ..., t_N(q_N)$, $1 \leq q_n \leq K (1 \leq n \leq N)$ can be obtained by minimizing the following objective function:

$$D(t_1(q_1), t_2(q_2), ..., t_N(q_N)) = \sum_{n=1}^{N} d_n(t_n(q_n), s_n) + \sum_{n=2}^{N} d_n(t_{n-1}(q_{n-1}), t_n(q_n)) \qquad (2)$$

where q_n is the candidate index of the shortlisted target segments corresponding the source segment s_n. The source-to-target distance $d_n(t_n(q_n), s_n)$ and boundary connectivity are both taken into consideration in the objective function.

A Viterbi algorithm [13] can be used to find the best segment sequence. Figure 1 illustrates the search of the optimal target segment sequence through the network, where $t_1(2), t_2(3), t_3(1), ..., t_N(3)$ segments are selected rather than $t_1(1), t_2(1), t_3(1), ..., t_N(1)$ segments which have the smallest source-to-target acoustic distance.

3. EXPERIMENTS AND RESULTS

3.1. Databases and experimental setups

The CMU ARCTIC database [14] is used to evaluate the performance of the proposed method. Two male (BDL, RMS) and one female (SLT) speakers are selected. 50 utterances of each speaker are selected for training, while 89 utterances, not included in the training set, are used as testing set. A male to male (BDL to RMS, M2M) and a male to female (BDL to SLT, M2F) conversion experiments are conducted.

The speech signal is sampled at 16k Hz and windowed by a 25ms Blackman window with a 5ms frame shift. 25 mel-cepstral coefficients (including the 0th coefficient) are extracted by applying a mel-cepstral analysis technique [15]. The fundamental frequency (F0) is automatically extracted on a short-time basis by using the robust algorithm for pitch tracking [16]. F0 is converted by equalizing the means and variances of the source and target log(F0) distributions: $x' = \frac{\sigma_y}{\sigma_x}(x - \mu_x) + \mu_y$ where μ_x and μ_y are the means and σ_x and σ_y are the standard deviations of the source and the target speakers, respectively; x is the source speaker's F0 and $x^{'}$ is the converted F0.

The BUT phoneme recognizer [17] is adopt to label the source and target speech and get the phoneme boundary.

In this work, we will compare the following four approaches.

- NNFA: the nearest neighbor frame alignment approach
- CNNFA: NNFA with phonetic label constraint
- NNSA: segment based alignment without the boundary connectivity constraint. As illustrated in Figure 1, the $t_1(1), t_2(1), t_3(1), ..., t_N(1)$ segment sequence.
- Proposed: segment-based frame alignment with target-to-target connectivity constraint.

Noted that the CNNFA method, which uses phonetic label as constraint, is similar to the class mapping based frame alignment [9]. NNFA and CNNFA are the two baseline approaches.

3.2. Objective evaluation

Mel-cepstral distortion (MCD) is used as an objective measure: $MCD = \frac{10}{\ln 10} * \sqrt{2 * \sum_{i=0}^{24}(mc_i^t - mc_i^c)^2}$, where mc^t and mc^c are the target and converted feature, respectively.

The percentage of correct phonetic alignment (CPA) [8] and correct phonetic class alignment (CCA) are also used as objective evaluation measures. CPA or CCA measures the percentage of how many frames are paired up with the frame with the same phoneme or phonetic class.

Table 2. Results for the four frame alignment approaches.

	M2M			M2F		
	MCD	CPA%	CCA%	MCD	CPA%	CCA%
NNFA	6.11	21.96	58.05	6.86	14.42	51.94
CNNFA	6.09	25.63	64.99	6.80	17.50	58.61
NNSA	6.10	32.94	66.76	6.61	24.28	62.08
Proposed	**5.91**	**34.69**	**67.04**	**6.44**	**24.87**	**61.34**

The two baseline results are presented in the first two rows in Table 2 as a reference for comparison with our proposed method.

To evaluate our proposed approach, we first check the effect of NNSA approach, segment based frame alignment without boundary connectivity constraint. From the results presented in Table 2, we find that the improvement for M2M conversion is limited when doing segment-based frame alignment, as the acoustic difference between males speakers is much smaller than that between male and female speakers.

Then we evaluate the performance of our proposed approach. Note that the cost of the search for best path with target-to-target connectivity distance, as shown in Figure 1, will increase as the size of the short list increase. A strategy is required to decide the number of shortlisted candidates. For male-to-male (M2M) segment selection, we find that the overall acoustic distance, which is the summation of source-to-target distance (ASD) and target-to-target connectivity distance (ABD), will converge when more than 10 candidate segments are pre-selected (See Figure 2); and from Figure 3, we find that the MCD between paired source-target frames of training is stable when 5 to 15 segments are shortlisted. Hence, we set the size of the short list to be 10 for M2M conversion. With the same reason, 13 candidates for M2F conversion are used. The MCD, CPA and CCA results of our proposed approach for both M2M and M2F conversion can be found in Table 2.

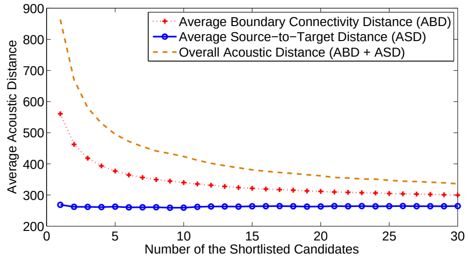

Fig. 2. *Overall acoustic distance (ASD + ABD) of the optimal target sequence with increasing the size of the short list.*

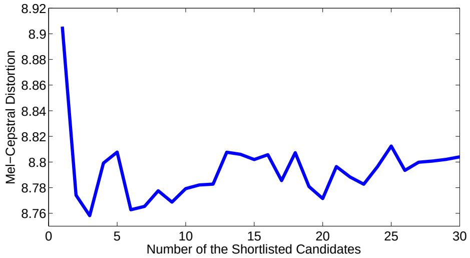

Fig. 3. *Mel-cepstral distortion (MCD) of paired source-target frames of training with increasing the size of the short list.*

In Table 2, comparing the results of our proposed approach with the two baseline approaches and the NNSA approach, significant improvement is observed for both M2M and M2F conversions. It shows that the target-to-target connectivity is important in finding a better frame alignment.

3.3. Subjective Evaluation

A number of ABX listening tests are conducted to compare our proposed method with the baseline approach CNNFA on M2M and M2F conversion (As the NNFA and CNNFA approaches have almost the same performance in objective evaluation). In each test, 8 listeners participated and 10 sentence pairs were used. The subjects first listened to the original target speech as a reference, then listened to the converted speech using two different methods. Subjects were asked to choose whether A or B sounded more similar to the target speech, or choose X when they could not hear difference. Results in Figure 4 consistently report the proposed method achieves better voice conversion performance than baseline approaches.

4. CONCLUSIONS

In this paper, we proposed a segment-based frame alignment method for text-independent and non-parallel voice conversion. In this approach, both the source-to-target acoustic distance and the target-to-target boundary connectivity distance are taken into consideration to find the best frame alignment. The experiments show the proposed methods achieved better performance than the conventional method.

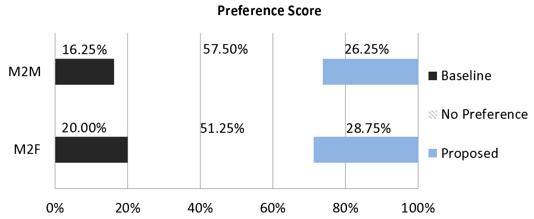

Fig. 4. *Results of subjective evaluation between CNNFA baseline approach and the proposed frame alignment .*

5. REFERENCES

[1] A. Kain and M. Macon, “Spectral voice conversion for text-to-speech synthesis,” in *ICASSP*, 1998.

[2] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[3] T. Toda, A. Black, and K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[4] A. Mouchtaris, J. Van der Spiegel, and 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.

[5] D. Sündermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, “Text-independent voice conversion based on unit selection,” in *ICASSP*, 2006.

[6] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Pérez, and Y. Stylianou, “Towards a voice conversion system based on frame selection,” in *ICASSP*, 2007.

[7] A. Black and P. Taylor, “Automatically clustering similar units for unit selection in speech synthesis.” in *Eurospeech*, 1997.

[8] D. Erro, A. Moreno, and A. Bonafonte, “Inca algorithm for training voice conversion systems from nonparallel corpora,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 944–953, 2010.

[9] D. Sündermann and A. Bonafonte, “A first step towards text-independent voice conversion,” in *ICSLP*, 2004.

[10] H. Ye and S. Young, “Voice conversion for unknown speakers,” in *ICSLP*, 2004.

[11] S. Siniscalchi and C. Lee, “A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition,” *Speech Communication*, vol. 51, no. 11, pp. 1139–1153, 2009.

[12] J. Hershey and P. Olsen, “Approximating the kullback leibler divergence between gaussian mixture models,” in *ICASSP*, 2007.

[13] L. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[14] J. Kominek and A. Black, “The cmu arctic speech databases,” in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[15] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech,” in *ICASSP*, 1992.

[16] D. Talkin, “A robust algorithm for pitch tracking (rapt),” *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[17] P. Schwarz, P. Matějka, and J. Černockỳ, “Towards lower error rates in phoneme recognition,” in *Text, Speech and Dialogue*. Springer, 2004, pp. 465–472.