International Journal of Speech Technology (2018) 21:79–84
https://doi.org/10.1007/s10772-018-9490-0

Improvement of time alignment of the speech signals to be used in voice conversion

Fatemeh Mozaffari[1] · Abolghasem Sayadian[1]

Received: 6 July 2017 / Accepted: 7 January 2018 / Published online: 15 January 2018

Abstract
One of the main applications of time alignment is parallel corpus based voice conversion. In the literature, various methods such as dynamic time warping (DTW) and hidden Markov model have been suggested for time alignment of two speech signals. In this paper, we introduce some modifications to DTW in order to decrease the time alignment error. These modifications are refinement, which is done by exerting a threshold, normalization, and comparisons between the preceding and the following frames to make sound correspondence between two different parallel corpus-based speakers' speeches. Evaluation of this approach which has been done on some corpus sentences indicates a significant improvement of time alignment. At least about 4% and in some cases 15% decrease of error in comparison with DTW has been achieved.

Keywords Dynamic time warping · Parallel corpus · Time alignment · Voice conversion

1 Introduction

Time alignment is holding sound correspondence between the frames of two speech signals. It is one of the important blocks in most applications of signal processing such as speech recognition, text to speech conversion and voice conversion (Yfantis et al. 1999; Rabiner and Juang 1993; Torkkola 1988; Homayounpour 2009; Sayadian and Mozaffari 2017). Time alignment is essential part of parallel corpus based voice conversion methods. To utilize the diverse methods of voice conversion which are trained through parallel corpus, we need to consider time alignment and sound correspondence among the different frames of signals of both the target and the source speakers' speeches (Tinati and Farhid 2007). This could lead to substitute each frame with its possible correspondent one regarding its sound. In this category of approaches, making sound correspondence has a remarkable effect on the effectiveness of voice conversion. The more accurate this correspondence is, the better the outcome results. Generally, the basic algorithms exploited for this purpose are DTW and HMM. Latsch and Sergio (2011) proposed a new joint time alignment and pitch modification algorithm by incorporating the pitch-synchronous characteristic onto the DTW alignment so, pitch modification and time alignment can be done simultaneously and it leads to a computationally efficient time-alignment process. Seara et al. (2016) used an automatic alignment tool by applying the speech recognizer for aligning each word and phoneme of the sentences which can be used to facilitate the completion of high-quality linguistic resources. In this paper, the basics of the suggested approaches are reviewed and then some solutions to improve the precision of time alignment presented in a novel framework. Finally, the performance and effectiveness of this approach are evaluated by using some sentences of a parallel corpus and in comparison to conventional methods.

2 DTW algorithm

One of the applications of DTW Algorithm is voice conversion. DTW is a method to hold sound correspondence between the frames of two speech signals of the same sentences (in the parallel case) despite different time length for each sound. The general approach in this regard is based on a series of local and global restrictions. DTW acts on the basis of the conformity of the articulated word with the

✉ Fatemeh Mozaffari
fa_mozaffari@aut.ac.ir

Abolghasem Sayadian
eeas35@aut.ac.ir

[1] Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran

pre-recorded patterns. In each comparison, the produced word is checked regarding its conformity with all the patterns. By using a cost function, we can evaluate the amount of this compliance thus the pattern with the minimum cost would be chosen as the recognized pattern (Wang and Cuperman 1998). In DTW, there are two distances to be computed: the local and the global ones. The local distance is the one calculated between the feature vector of one signal and that of another one. Supposing that **Y** is the feature vector of one frame of the first signal pattern and **X** is that of the second signal of the unknown word, we can come up with the distance between these two frames via formula (1) (Wang and Cuperman 1998).

$$\mathrm{d}(\mathbf{X},\mathbf{Y})\sqrt{\sum_{n=1}^{N}(\boldsymbol{X}_n^i-\boldsymbol{Y}_n^j)^2}\,, \tag{1}$$

Bold and upper case notation used to indicate the vectors. i and j are the indexes of the frames related to each of the two signals, and *N* is the indicator of the number of features.

The mentioned function is appropriate to all is and js which leads to create a matrix of distances. Then we start from the first element of the matrix and reach the final one through the most reasonable one i.e., with the minimum distortion. Then the following formula determines the global distance to the point (i,j); e.g., for the restriction type1 we have (Wang and Cuperman 1998),

$$\mathrm{D}(i,j)=\min\{D(i-1,j-1),D(i,j-1),D(i-1,j)\}+d(i,j). \tag{2}$$

This distance has already been used in the comparison of the unknown word with each of these patterns, and the pattern causing the minimum cost or minimum global distance is selected as the recognized word. Figure 1 shows some samples of the local restrictions which are common in conventional DTW.

As mentioned before, there are some local and global restrictions to apply to the search space, based on which the maximum and the minimum extension in one direction can be determined. Formulas (3) and (4) indicate the upper and lower bounds of the global restrictions (Wang and Cuperman 1998).

$$1+\frac{[\varphi_x(k)-1]}{Q_{\max}}\le\varphi_y(l)\le 1+Q_{\max}\left[\varphi_x(k)-1\right] \tag{3}$$

$$T_y+Q_{\max}\left[\varphi_x(k)-T_x\right]\le\varphi_y(k)\le T_y+\frac{[\varphi_x(k)-T_x]}{Q_{\max}}\,. \tag{4}$$

Therefore, DTW Algorithm can be summarized in three steps being presented in formulas (5–7) (Wang and Cuperman 1998).

1. Initialization

$$D_A(1,1)=d(1,1)m(1), \tag{5}$$

2. Recursion

$$D_A\left(i_x,i_y\right)=\min_{\left(i'_x,i'_y\right)}[D_A(i_x{}',i_y{}')+\xi((i_x{}',i_y{}'),(i_x,i_y))], \tag{6}$$

3. Termination

$$d(\boldsymbol{X},\boldsymbol{Y})=\frac{D_A\left(T_x,T_y\right)}{M_\phi}. \tag{7}$$

$\xi((i_x{}',i_y{}'),(i_x,i_y))$ represents each path spot of the zone from which we are allowed to move and reach $\left(i_x,i_y\right)$. So,

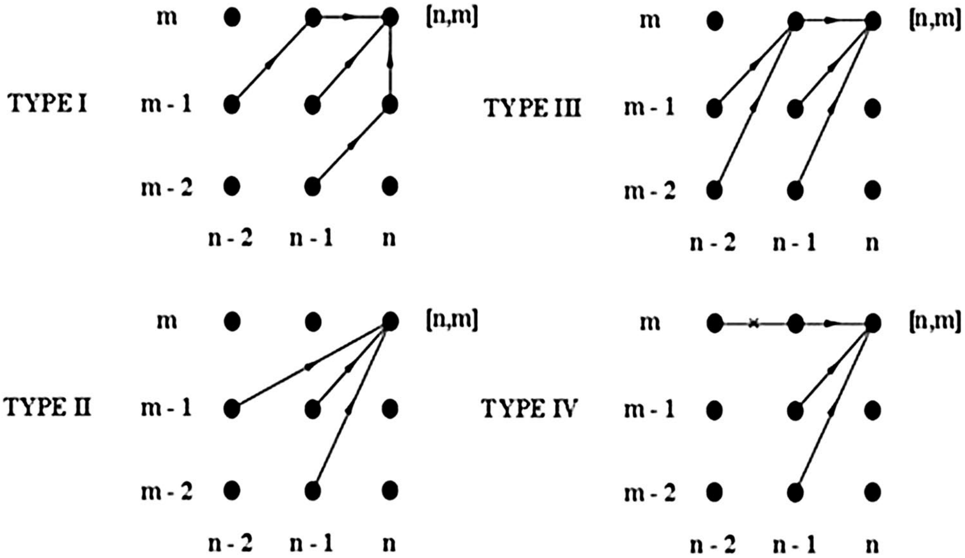

Fig. 1 Local restrictions used in DTW algorithm

by d(**X, Y**) we mean the distortion between **X** and **Y** signals. We should apply the global restriction of the path while applying the Algorithm, and any deviation from the parallelogram zone, which is being determined by formulas (3) and (4), is not permitted (Wang and Cuperman 1998; Stainhaouer and Carayannis 1990).

3 Using HMM for time alignment

One of the other approaches used in time alignment and creation of sound correspondence between the frames of two speech signals is using HMM. Here it is assumed that we have the identical sentences from two speakers and then we apply HMM Algorithm. This Algorithm is one of the smart approaches to classify the patterns (Rabiner 1989). HMM models the temporal and positional alterations of the statistical characteristics of a random process through a Markov Chain which has state-oriented stationary sub-processes; in other words, an HMM is basically a dynamic Bayesian network with restrictive conditions which model the transmissions between the states. Moreover, a set of probability density functions model the random changes of the signals in each state. Generally an HMM has *N* states each of which is applicable to model a separate part of a signal process.

Basically, there are two main changes considered in speech signals and other random ones including changes in spectral combination and changes in the temporal scale or speech rate. In an HMM, these changes are modeled via probability of transmission, and observation of states (Rabiner 1989). The useful method to interpret and apply the HMM methods is considering each state of the HMM as a model of a part of the random process. In order to implement the HMM, Viterbi Method and Baum-Welch Algorithm are used (Arslan and Talkin 1998). In this paper, the sentence HMM to align the frames of two speech signals is elaborated on. The Algorithm to hold the sound correspondence which can be used in voice conversion is as follows (Dengï and Byrne 2008).

1. The same sentences from two speakers are considered.
2. The silence parts in each speech signal are omitted.
3. For each frame, Cepstral Coefficients, Delta Cepstrum, Logarithm of energy, and whether it is voiced or unvoiced are determined.
4. The mean values are subtracted in order to gain robust spectral estimation.
5. Based on the sequence of the parameter vectors, sentence HMMs are trained for each target speaker's speech utterance.
    5.1. The number of states in sentence HMM pertinent to temporal duration of each speech utterance is determined.
6. Instruction is accompanied by applying Baum-Welch Algorithm.
    6.1. In the primary instruction, to prevent the unnecessary states, the spectral vectors with similar mean, are considered as a unified state.
    6.2. The initial covariance matrix is estimated due to training data set.
7. The best state sequence for each speech utterance is measured using Viterbi Algorithm.
8. The LSF mean vector for each state of both the target and source speakers is measured by using the frames associate with the state index.
9. The LSF mean vectors are accumulated for each and every sentence to arrange the codebook of the source and target speakers.

4 A proposed approach to improve the alignment

Considering the fact that the data relevant to the speaker's identity are in the voiced frames, we can omit the unvoiced ones to perform the alignment process to be used in voice conversion. Then with the help of the Euclidean distance comparison between the adjacent feature vectors, we can prevent the repetition of the vectors the distance of which is less than the supposed threshold; i.e., we perform a sort of refinement on each speaker's feature vectors so that very portion of temporal difference between two speeches, which is usually due to the length of vowels being pronounced, fades away. Then to equalize the temporal length of the two speeches, we count on the bigger length as the criterion rather than the smaller one. So we do not waste any data by lengthen the shorter temporal length with the linear interpolation. Utilizing normalization based on the mean and standard deviation could be considered as an approach to improve the algorithm, especially to hold time alignment between two speakers, which is our purpose in voice conversion; i.e., we calculate the mean and standard deviation for all of the frames and for each element of the feature vector. The next step is conducting normalization by subtracting the mean and dividing the outcome by standard deviation. Through all these, the feature spaces of the two speakers become more analogous, and the sound correspondence will take place much more flawlessly. In DTW, particularly in its typical mode, the paths to hold this correspondence between the two speeches, submits to certain local and global restrictions in none of which is it feasible to make a comparison between a frame and its preceding ones of the second speech, as the direction is always forwarding. Now in case the frame correspondence is recognized incorrectly, the feasibility of comparison for the next frame with their previous ones fades, that is, the possibility of holding a correct correspondence will be lost. That is why another point to be taken into account is that we should provide each frame with the feasibility of being compared with corresponding ones whether before or after. The number of these frames is dependent on the length of the two speeches (becoming identical after interpolation). So the instruction of the proposed algorithm is as follows:

1. The 26-feature vectors (Cepstrum Coefficients, Delta Cepstrum, the Normalized Energy and Delta Energy) are made for each 8 ms frame of the two speakers' signals.
2. After defining the beginning and the ending of these speech files* for the both speakers, we omit the frames labeled as silence or unvoiced.
3. Following obtaining the spectrum distortion between each two adjacent frames, we can omit the next frame if the obtained distortion is less than the proposed threshold, so we can continue with the next one.
4. Linear interpolation is exerted on the speech signal which has fewer frames (i.e., of shorter length) so as to hold sameness between the numbers of the frames of the two speeches to be aligned.
5. Pursuing normalization would be through subtracting the mean and dividing it over the standard deviation to gain more similarity between the spaces of the two speeches.
6. Saving the distortion (Euclidean distance) between the feature vectors of the two speeches in a matrix is the next step.
7. Setting one of the speeches as reference is so important, as each of its frame is going to be compared with the frame of the same index of another speech and the αT of the previous and the following frames. T refers to the number of the frames of each speech, and α is considered between 0.05 and 0.3 which is obtained by empirical method. The frame with the least distortion (distance) in comparison to the target frame is chosen.

So, the main modifications to DTW algorithms are:

- Since the information about speaker's identity is in voiced frames, unvoiced frames have been eliminated for alignment to be used in voice conversion.
- Refining the frames by eliminating the adjacent frames which have the distortion less than the desired threshold between them. So that, the time difference between two signals due to the extension of vowel utterance would be ignored.
- Using interpolation to hold the sameness between the numbers of frames of two signals.
- Normalization to mean and standard deviation which improves the accuracy of alignment.
- Comparison between frames is done by comparing each frame to the frames before and after the corresponding frame from the other signal. The number of frames is determined by a threshold. In this case, if we made a mistake in finding a corresponding frame for one vowel, it can be modified in the next comparison by comparing to the previous frames.

5 Start point and end point detection

The automatic definition of the beginning and the ending of each speech file have been made based on its periodic entity using the normalized cross-correlation and the gain. Following this process would be finding the first frame the periodic measure of which is more or equal to 0.8 (the desired amount of this measure for voiced frames) along with the gain of more than or equal to 52 dB (the gain for voiced frames) at the beginning of each speech file. Worthy to mention is that the accurate measure of being periodic and the gain for being diagnosed as voiced would be conducted by performing a test on different sentences and acquiring an amount representing the correct result. These definite measures have been the outcome of this research regarding the exercised data. Providing that these features are considered true about the next 6–8 frames, this very frame will be acknowledged as the first voiced frame in a speech file. Then the previous frames going back to 25 ones (as long as 8 ms) are entitled as unvoiced frames before the first frame which is added to the feature matrix of the mentioned speeches if and only if they are not classified as silence; otherwise, the sentence must have been started with a voiced phoneme, and we manage to define the beginning of the speech as well as the ending benefiting the same method.

6 Evaluation

To analyze this algorithm, optimum values for two parameters have been obtained by changing them. These parameters are threshold and α. Tables 1 and 2 indicate the results. We have also used DTW with one of the local restrictions the error of which is the least for our experiment so that we can compare the proposed algorithm to conventional DTW by comparing errors based on Euclidean distance between frames of two signals. In Tables 1 and 2, DTW Type refers to the type of local restriction used to hold time alignment in DTW. The frames for comparison due to the suggested approach are those which are compared before and after each

Table 1 Holding sound correspondence between the speech signals of two male speakers

Threshold	α	DTW	The number of frames for comparison	The total number of frames	Error (%)
0.004	0.09	–	6	62	50
0.005	**0.01**	–	**1**	**58**	**12**
0.005	0.02	–	2	58	13.7
0.005	0.05	–	3	58	17
0.005		Type1	–	58	14
0.005		Type2	–	58	14
0.006	0.05	–	3	50	33

Bold values show the conditions i.e. threshold and α by which the error of the proposed algorithm becomes minimum

Table 2 Holding sound correspondence between the speech signals of a male and a female speakers

Threshold	α	DTW	The number of frames for comparison	The total number of frames	Error (%)
0.004	0.01	–	1	62	24
0.005	**0.01**	–	**1**	**58**	**12.87**
0.005	0.02	–	2	58	18.6
0.005	0.04	–	3	58	23.8
0.005	0.06	–	4	58	31
0.005	0.1	–	6	58	40
0.005	–	Type2	–	58	36
0.006	0.06	–	4	53	32

Bold values show the conditions i.e. threshold and α by which the error of the proposed algorithm becomes minimum

frame from one speech with those of another one. For example, if there are 3 frames to compare, the nth frame from the first speech is compared to the ones from n − 3 to n + 3 of the second speech, and the frame which leads to minimum distortion is chosen as the recognized one.

As mentioned before, α is a co efficient which determines the number of frames to be compared. This number is obtained by multiplying the total number of frames of the two speeches by α. The threshold addressed in these tables is the criterion upon which two consecutive frames are compared to each other in terms of distortion. Therefore, the frame which has the less-than-threshold distortion with reference one would be omitted.

Noteworthy to point is that the total number of frames is the number the frames of the two speeches have reached after refinement and interpolation.

The error amount has been determined by dividing the number of the mistakenly corresponded frames over the total number of frames. As observed, the amount of error depends on two parameters. One of them is the threshold used to compare the adjacent frames with each other in order to refinement, and the other is parameter α which determines the number of frames for comparison in order to find the one with minimum Euclidean distance.

The results of the evaluation indicate that the optimum determination of these two parameters, especially α, is different for various speeches and depends on some factors such as the length difference of the two speeches to be corresponded. Of course, the results have generally demonstrated that the threshold from 0.004 to 0.006 provides an appropriate number of frames for each phoneme; that is why the threshold in these tables has been considered 0.005. Another observed result was that the closer the two frames of the two speeches after refinement which leads to fewer required frames for comparison, the fewer the errors.

In the next step, four speakers (two females and two males), and 20 sentences for each have been used. After identifying the beginning and the end of the speech frames for each speaker, the ones titled as silence or unvoiced are omitted, then first, normalization process for the gain through subtracting the maximum gain from the other gains, and second, obtaining the mean and the standard deviation for each of the features from the 26-character feature vector are performed. What is done to normalize all the elements of the feature matrix is to subtract the mean and dividing it over the standard deviation. Noticeable is that normalization is done after refinement. Finally, utilizing comparison of α times of the total number of frames before and after each frame, we can find the corresponding frame with that frame. As follows, Table 3 focuses on the mean and the standard deviations for one male and one female, two male and two female speakers; also, the number of frames to compare has been achieved by multiplying α by the total number of frames. This method leads to the fewest errors of time alignment regarding the data.

in this research. For instance, for the first sentence with 80 frames, if we consider α equal to 0.04 so compare each frame to the same index frame and also three frames before and after that frame, and threshold equal to 0.005, the errors of time alignment are fewer than the number of errors resulted in other approaches; i.e., this number of frames to

Table 3 Mean and the standard deviation of time alignment through the suggested approach for 20 sentences and different combinations of speakers

Combination of speakers	Mean of error (%)	Standard deviation of error
M–M	14	6%
F–F	13.58	4.61
M–F	15.83	3.63%

compare is optimum value for sentence number 1. The same goes for other sentences being tested by the same approach to acquire the optimum value. Finally, the mean and the standard deviation of the obtained errors are calculated for 20 sentences. Table 3 provides the results.

7 Conclusion and suggestions

As mentioned, time alignment is one of the most significant steps in the process of voice conversion based on parallel corpus. In a predominant approach to hold sound correspondence, DTW is used from predetermined paths based on the local and global restrictions and ranking functions applied to reach the least distortion among the frames. The method suggested in this article relies on the general idea of comparison between the frames of one speech to another one in order to choose the frame that has minimum Euclidean distance from reference frame, being proposed in DTW. The distinct instructions of our approach are: refinement based on the threshold, normalization, and comparison not only to the same index frame but also to the preceding and following frames to hold the sound correspondence between two different speakers' speeches which are based on parallel corpus. This very evaluation on some verbal corpus sentences demonstrated that errors decrease considerably (at least 4%, and in some cases about 15%) in comparison to those obtained by using DTW. Noteworthy to mention is that the desired threshold to attain the minimum error and the number of the frames being compared to one another on each spot of the path are considered as two independent parameters which could be optimized for multifarious cases. To achieve a high level of improvement of time alignment, it is recommended to use vowel recognition and repeated vector quantization. Moreover, as we can see in Table 3, mean of error is higher for combined speakers. So we should focus on a different approach for combined speakers in order to improve the method further.

References

Arslan, L. M., & Talkin, D. (1998). Speaker transformation using sentence HMM based alignments and detailed prosody modification. *ICASSP*.

Dengï, Y., & Byrne, W. (2008). HMM word and phrase alignment for statistical machine translation. *IEEE Transactions on Audio, Speech and Language Processing, 16*, 494–507.

Homayounpour, M. (2009) *Text to speech conversion*. Tehran: Amirkabir University of Technology.

Latsch, V. L., & Sergio, L. N. (2011). Pitch-synchronous time alignment of speech signals for prosody transplantation. *IEEE international symposium on circuits and systems (ISCAS)*.

Rabiner, L., & Juang, B. H. (1993). *Fundamentals of Speech Recognition*. Upper Saddle: Prentice Hall.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech. *Proceedings of the IEEE*.

Sayadian, A., & Mozaffari, F. (2017). A novel method for voice conversion based on non-parallel corpus. *International Journal of Speech Technology*. https://doi.org/10.1007/s10772-017-9430-4

Seara, R., et al. (2016). Enhanced CORILGA: introducing the automatic phonetic alignment tool for continuous speech. *LREC*.

Stainhaouer, G. N., & Carayannis, G. (1990). New parallel implementations for DTW algorithms. *IEEE Transactions on Acoustics Speech Signal Processing, 38*, 4.

Tinati, M., & Farhid, M. (2007) A novel method for improvement of the quality of voice conversion systems. *13th national computer engineering conference of Iran*.

Torkkola, K. (1988). Automatic alignment of speech with phonetic transcriptions in real time. *Proceedings of IEEE*.

Wang, T., & Cuperman, V. (1998). Robust voicing estimation with dynamic time warping. *Proceedings of IEEE.*.

Yfantis, E. A., Lazarakis, T., & Angelopoulos, A. (1998). On time alignment and metric algorithms for speech recognition. *Proceedings of IEEE*.