# Segmentation of Speech Signals in Template-based Speech to Singing Conversion

Ling CEN, Minghui DONG, Paul CHAN
Institute for Infocomm Research (I2R), A*STAR, 1 Fusionopolis Way, Singapore 138632
E-mail: {lcen, mhdong, ychan}@i2r.a-star.edu.sg Tel: +65-963756291

***Abstract*— Singing voice synthesis has found numerous applications in the entertainment industry over the recent years. The template-based personalized singing voice synthesis method is a new method of generating high quality singing voice, which synthesizes the singing voice by means of conversion from the narrated lyrics of a song. In this synthesis method, template speaking and singing voices are first recorded for the purpose of modeling the transformation from speech to singing. To improve its accuracy while reducing computational load, the template voices are divided into several segments so that fine alignment and subsequent conversion can be performed separately for each segment. To correctly generate singing voice, a new instance of speech has to be divided into similar segments, each containing the same stanza as in the template voices. In order to achieve this, an automatic segmentation method is proposed in this paper. The experiment results have shown that the segmentation of speech signals using our method is comparable to manual segmentation, with an accuracy of 98.24%. This performance is consistent even in the presence of noise.**

## I. INTRODUCTION

With the development of the computer-based music technology, there has been a growing interest in singing voice synthesis over the recent years [1-6]. Former techniques in [1-3] are mostly focused on singing voice synthesis from the lyrics of a song, which is called Lyrics-to-Singing (LTS) synthesis. Generating singing voices from spoken lyrics is called Speech-to-Singing (STS) synthesis [4-6]. Compared to LTS synthesis, the advantage of STS synthesis is that the timbre of the speaker can be preserved and the singing can, consequently, sound like it is being sung by the original speaker. STS synthesis can have many applications in the view of practice since it enables the user to synthesize and listen to his/her own singing voices by simply reading the lyrics of songs.

In STS methods, conversion of speech into singing can be accomplished with either the guidance of music score or by means of transformation models based on pre-recorded templates of the song lyrics being read and sung synchronously. Unlike the score-based STS synthesis method in [4-6] that modeling the conversion of acoustic features based on the music score of a song, template-based STS synthesis trains the transformation models by analyzing a pair of template speaking and singing voices that are usually from a professional singer. Compared to score-based STS synthesis, it has two main advantages. Firstly, it omits the need to input the music score, which simplifies the operation of the system. Secondly, the pitch contour is derived from the actual singing voice, which is more natural than modifying a step contour to account for pitch fluctuations such as overshoot and vibrato. This can potentially improve the naturalness and quality of the synthesized singing.

In template-based STS synthesis, transformation models of acoustic features are trained by analyzing the template speaking and singing voices. The lyrics of a song usually have multiple phrases. In view of accuracy and computational load, template voices are necessarily divided into several segments. The transformation models are trained separately for each segment. To apply these models to the conversion of a new speaking voice, it has to be first divided into similar segments, each containing the same stanza as the template voices. Without accurate segmentation, the singing voice cannot be generated correctly.

Automatic segmentation of speech signals, however, is a very challenging task, which has been investigated in many speech research areas, e.g. Automatic Speech Recognition (ASR) [7-11]. The segment boundaries can be determined by inspection of speech waveforms and spectrograms based on the difference between the features of speech and silence [7]. However, these methods are not robust enough to divide the speech signal into the desired segmentations accurately all the time. This happens because factors such as signal-to-noise-ratio (SNR), speaking rate, pause duration and unexpected pauses within speech, may affect the segmentation and change the total number of segments and the number of words in one segment. Recently, approaches such as Hidden Markov model (HMM), Maximum Entropy (Maent) and Conditional Random Field (CRF) classifiers have been applied and combine both lexical and prosodic features [8-11]. Although they have been shown to have better accuracy than a pause-based segmentation, the process is much more complicated by formulating boundary-event detection as a sequence tagging problem, where each word in the speech has to be assigned a boundary label to the interval between that word and the next. Furthermore, these methods are applied after speech recognition and the information about alignment of word and phone transcriptions with the acoustic speech signal is required in feature extraction and segmentation [8].

To address this problem, a Dynamic Time Warping (DTW) based segmentation method, together with a method for silence removal, is proposed in this paper. With the help of our method, the speech signal can be segmented consistently with the template. Besides high accuracy in automatic segmentation, easy implementation and simple process are also the advantage of our method.

The remaining parts of this paper are organized as follows. The template-based STS system is briefly introduced in Section II, in which, the segmentation method will be applied. Following that, the methods for the silence removal and DTW based segmentation are elaborated in Sections III and IV, respectively. The experimental results are presented in Section V. The concluding remarks are given in Section VI.

## II. Segmentation in Template-based STS Synthesis

The STS system aims to convert a speaking voice into singing by automatically modifying the acoustic features of the speech. The template-based STS synthesis system may be broken down into three stages, namely, the learning, transformation and synthesis stages. In the learning stage, the template singing and speaking voices are analyzed to derive the transformation models for the conversion of speaking to singing based on the modification of acoustic features such as pitch contour, phoneme duration and spectrum. In the transformation stage, features are extracted for the speaking voice to be converted which is usually uttered by a different person. These are modified to approximate those of the singing voice based on the transformation models. After these features have been modified, the singing voice is synthesized in the last stage.

As described in Section I, the template voices have been manually segmented and the transformation models are derived separately for each segment. To convert a new instance of speech into singing using the trained transformation models, it has to be segmented similarly to the template speech. Inaccurate segmentation leads to incorrect transformation and consequently produces incorrect singing voices.

To address this problem, a segmentation method is proposed, whose flowchart is shown in Fig. 1. First, the speech is processed by removing the silent frames from the signal for the purpose of accurate alignment. Next, it is aligned with the template speech that has been manually segmented prior to conversion. The segment boundaries are then derived according to the synchronous information. The details are elaborated in the following two sections.

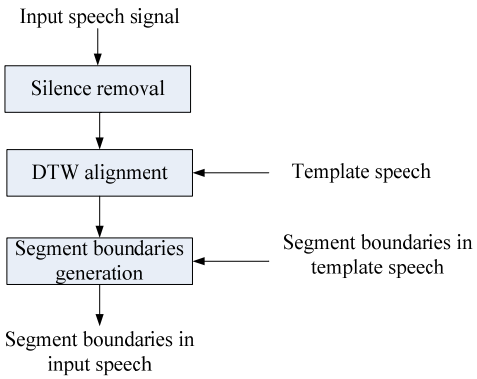


Fig. 1 Flowchart of the proposed method for speech segmentation.

## III. Silence Removal

To remove the silence from a speech signal, two features, i.e. short-time energy and spectral centroid, are extracted from the speech signal. Since it is assumed that the signal within a short period is stationary or quasi-stationary, the signal is divided into frames, each of which has a length of 20 msec.

Let $s_i(n), n = 1,2,...,N$ be the audio samples in the $i^{\text{th}}$ frame and its short-term energy, denoted as $e_i$, can be expressed as

$$e_i = \frac{1}{N}\sum_{n=1}^{N}\left|s_i(n)\right|^2, \tag{1}$$

where $N$ is the number of samples in one frame. Energy is the most useful feature to discriminate silence from speech in speech signals.

The spectral centroid can be defined as

$$c_i = \frac{\sum_{k=1}^{K}(k+1)S_i(k)}{\sum_{k=1}^{K}S_i(k)}, \tag{2}$$

where $S_i(k), k = 1,...,K,$ is the Discrete Fourier Transform (DFT) coefficients of $s_i$.

Let the first two local maxima of the histogram of the energy sequence be $E_1$ and $E_2$, and those of the spectral centroid be $C_1$ and $C_2$. The thresholds used to discriminate the speech and silence frames are defined as [7]

$$T_e = \frac{W_e \times E_1 + E_2}{W_e + 1}, \tag{3}$$

and

$$T_c = \frac{W_c \times C_1 + C_2}{W_c + 1}, \tag{4}$$

where $W_e$ and $W_c$ are the weighting coefficients for $T_e$ and $T_c$, respectively. Instead of using a fixed value in [7] as

$$W_e = W_c = 5, \tag{5}$$

the weights are defined as

$$W_e = 0.4 \times E_2 / E_1, \tag{6}$$

and

$$W_c = 0.5 \times C_2 / C_1. \tag{7}$$

This is more reasonable as the threshold need to be adjusted based on the level of the input signal.

The thresholds defined in (3) and (4) are used to detect the start of speech activity. Let $s_{sl}(n), sl = 1,2,...,SL$ be the first $SL$ frames absent of the speech and the speech activity starts from $(SL+1)$-th frame.

Knowing the starting point of the speech activity, we re-define the thresholds of energy and spectral centroid as

$$T_e = \frac{\sum_{sl=1}^{SL} e_{sl} + 0.1 \times \min_{sp=1}^{SP}(e_{sp})}{SL+1}, \tag{8}$$

and

$$T_c = \frac{\sum_{sl=1}^{SL} c_{sl} + 0.1 \times \min_{sp=1}^{SP}(c_{sp})}{SL+1}, \quad (9)$$

where $e_{sl}$ and $c_{sl}$ are the energy and spectral centroid of $s_{sl}(n)$, respectively, $e_{sp}$ and $c_{sp}$ are the energy and spectral centroid of speech frames, respectively, and $SP$ is the total number of speech frames. The speech segments are formed by successive frames with features values larger than the thresholds. The silence is removed from the speech signal and the resultant signal will be segmented using the method described in the following section.

## IV. Speech Segmentation

After the silence has been removed from the speech signal, segmentation is carried out. Let $s_S(l), l = 1,2,...,L_S$ be the speech signal under segmentation, and $s_T(l), l = 1,2,...,L_T$ be the template speech signal, which have $L_S$ and $L_T$ audio samples, respectively. Here, the template speech, $s_T$ has been manually segmented to have $G_T$ segments, each of which is denoted as

$$s_{Tg}(l), g = 1,2,...,G_T.$$

Let $s_{Tg}(l_{Tg})$ be the last sample of the $g^{\text{th}}$ segment, where $l_{Tg}$ is its position.

The boundary samples of $s_S$, which is represented as $s_{Sg}(l_{Sg})$, are found by aligning $s_S$ and $s_T$. Alignment is performed using DTW [12]. DTW compares the similarity between each frame of a sequence with every frame of another, which may vary in time, seeking to match the corresponding frames between them. This method has been largely used in ASR to deal with different speaking speeds.

The short-time cepstral features, MFCC, are extracted as acoustic features in alignment. Besides the MFCC, the Delta and Acceleration (Delta-Delta) of the raw MFCC features are calculated. There are 39 MFCC features in the full feature set, including 12 MFCC features, 12 delta MFCC features, 12 Delta-Delta MFCC features, 1 (log) frame energy, 1 Delta (log) frame energy, and 1 Delta-Delta (log) frame energy. In order to reduce the acoustic variation across different frames and different parameters, frame- and parameter-level normalizations are carried on the MFCC features. Normalization is performed by subtracting the mean and dividing by the standard deviation of the features.

With the help of alignment, the synchronous information between $s_T$ and $s_S$ can be obtained. As long as we know $s_{Tg}(l_{Tg})$ of $s_T$, we can estimate $s_{Sg}(l_{Sg})$ for $s_S$ from the synchronous information. In this way, the segmentation of $s_S$ can be always consistent with that in $s_T$ regardless of other factors, e.g. SNR and pause duration, etc.

## V. Evaluation

To evaluate the performance of the proposed method, a popular Chinese song titled “why do you bear to hurt me so” was selected in the experiment. Two singers, one male and one female, were employed to read the lyrics of the song to achieve 2 spoken utterances which were used as the templates for both genders. Each template utterance having around recording of 50-60 seconds was manually divided into 17 segments.

The data used for testing were 10 spoken utterances that were read by 6 speakers including 2 males and 4 females. Each utterance read the same lyrics as the template speech. For the purpose of comparison, the utterances were segmented using 4 methods listed as below:

A. Proposed method including silence removal in Section III and DTW based segmentation in Section IV
B. Silence removal in [7] and DTW based segmentation in Section IV
C. DTW based segmentation in Section IV only
D. Method proposed in [7].

The results are represented using the error rate and error types. The error rate (ER) is defined as

$$ER = \frac{Seg_{wrg}}{Seg \times Utt}, \quad (10)$$

where $Seg_{wrg}$ is the number of segmentation errors, $Seg$ is the number of segments in one utterance, and $Utt$ is the number of utterances in testing. The error types (ET) are listed below

a) Grouping multiple segments into one segment
b) Dividing one segment into multiple segments
c) Missing starting- or ending- words in one segment
d) Having only silence in a segment
e) Early segmentation (Part of the last syllable in a segment is carried over to the subsequent segment)
f) Late segmentation (Part of the first syllable in a segment is segmented with the previous segment)

The 10 spoken utterances were segmented using the 4 methods. For each method, the average value of the error rates and the number of errors for each type in the 10 utterances were tabulated as shown in Table I.

It can be seen from Table I that our method (Method A) has the lowest ER of 1.76% among the 4 methods. There are only 3 errors in dividing the 10 utterances into 170 segments.

TABLE I
SEGMENTATION RESULTS

| Method | ER (%) | ET (error times) | | | | | |
|---|---|---|---|---|---|---|---|
| | | a | b | c | d | e | f |
| A | 1.76 | | | | | | 3 |
| B | 7.06 | | | 2 | | 8 | 2 |
| C | 4.12 | 1 | | | 1 | | 5 |
| D | 30.59 | 23 | 17 | 6 | 6 | | |

Taking one spoken utterance as an example, the segmentation results are shown in Figs 2 and 3. The optimal warping path (red line) in the time warping matrix, when this utterance is aligned with the template speech, is illustrated in Fig. 2. The blue circles in this figure show the segment boundaries. The waveforms of the 17 segments are illustrated in Fig. 3(a). The adjacent segments are discriminated using different colours, either green or red. It can be seen from Fig. 3(a) that a noise (in blue) with much larger amplitude than that of the speech appeared at the end of signal, which was not recognized as a part of the last segment but filtered out by our method.

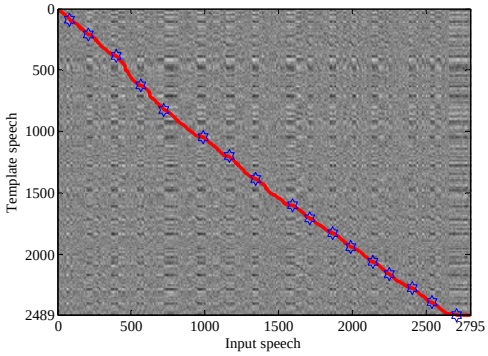


Fig. 2 Alignment of the input speech and template speech.

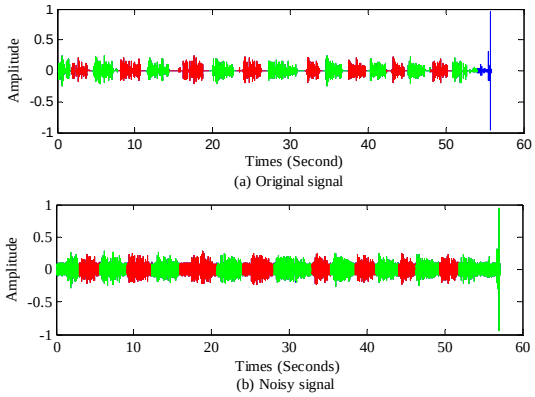


Fig. 3 Waveforms of the 17 speech segments in the original signal (a) and the noisy signal with white Gaussian noise (SNR = 2) (b).

To evaluate the performance of our method in noisy condition, white Gaussian noise was added to the signal given in Fig. 3(a) with a low SNR of 2. The noisy signal was correctly divided into 17 segments without any error by using the proposed method, whose waveforms are illustrated in Fig. 3(b). It shows that this method is able to accurately find the boundaries of the segments even when there is large noise in the speech signal. This is further indicates the robustness of our method.

It can be found from the experiment that the removal of the silence from both the segmented and the template speech is critical in improving the performance of segmentation. If the original speech signals were segmented together with the silence as described in Method C, the ER was increased to 4.12%. However, if the silence was removed using the method proposed in [7], ER reached 7.06% (see Method B). When the segmentation method in [7] was employed, the error rate is 30.59%, which is much higher than that achieved by our method.

## VI. Conclusions

In template-based personalized singing voice synthesis, the input speech has to be divided into several segments before its features can be converted using the transformation models. Its segmentation should be consistent with that of the template voice. To achieve this, a segmentation method is proposed in this paper. First, silence is removed from the speech signal. Then, it is aligned with the template speech using DTW. Corresponding segment boundaries are derived with the alignment information attained. The results of the experiment show that the segmentation error can be as low as 1.76% even in the presence of noise.

## References


[1] Bonada, J. and Serra, X., "Synthesis of the singing voice by performance sampling and spectral models," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 67-79, March, 2007.

[2] YAMAHA Corporation, Vocaloid: New singing synthesis technology, *http://www.vocaloid.com/en/index.html*.

[3] Saino, K., Zen, H., Nankaku, Y., Lee, A., and Tokuda, K., "HMM-based singing voice synthesis system," in *Proc. Interspeech*, pp.1141-1144, Sept. 2006.

[4] Saitou, T., Goto, M., Unoki, M., and Akagi, M., "Speech-to-Sing synthesis: converting speaking voices to singing voices by controlling acoustic features unique to singing voices," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 215-218, Oct. 2007, New Paltz, NY.

[5] Saitou, T., Tsuji, N., Unoki, M., Akagi, M., "Analysis of acoustic features affecting "singing-ness" and its application to singing-voice synthesis from speaking-voice," in *Proc. Interspeech*, vol. 3, pp. 1929-1932, 2004.

[6] Ota, K. and Ehara, T., "Four-tone modeling for natural singing synthesis in Chinese and comparing synthesized singings with speaking voices," in *Proc. International Congress on Acoustics*, pp. 1-4, August 2010, Sydney, Australia.

[7] Giannakopoulos, T., "A method for silence removal and segmentation of speech signals, implemented in Matlab," Report presented in Mathworks File Exchange Website, 2010.

[8] Ostendorf, M., Favre, B., Grishman, R., Hakkani-Tür, D., Harper, M., Hillard, D., et al., "Speech Segmentation and Spoken Document Processing," *IEEE Signal Processing Magazine*, pp. 59-69, May 2008.

[9] M. Tomalin and P. Woodland, "Discriminatively trained Gaussian mixture models for sentence boundary detection," in *Proc. ICASSP*, 2006, pp. 549–552.

[10] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," in *Proc. ICSLP*, 2002, pp. 917–920.

[11] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[12] Sakoe, H. and Chiba, S., "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43- 49, February 1978.