PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

A simple strategy for natural Mandarin spoken word stretching via the vocoder

Yi-Jhe Lee*, Ting-Chun Liao, Yi-Wen Liu

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan,

Abstract

In Mandarin, when a word is spoken with a longer duration, different regions of it are not stretched uniformly. Moreover, transition between consonants and vowels may be hard to define. Therefore, it is challenging to find the stretchable regions of words automatically. In this paper, we explore the idea of parsing a Mandarin word into the part that should be played with the original speed followed by a uniformly-stretchable region. Then, the optimal dividing point to start stretching could be determined automatically by minimizing the distance between the stretched version (generated by computer) and a ground truth (spoken by human). A database of 42 pairs of regular-speed and slow utterances were created. The dividing points on the regular-speed utterances were found as proposed. The points could be aligned to words with the same pronunciations in full sentences by dynamic time warping, and the sentences could be synthesized with arbitrary tempo and rhythms. The naturalness of stretching by three methods was evaluated subjectively: uniform stretching by waveform similarity-based overlap-add (WSOLA), uniform stretching based on linear interpolation in the vocoder domain (LI-VD), and the proposed strategy. 79.6% of answers chosen by 41 subjects show that our method outperforms.

Keywords: Speech, time-scale modification, Mandarin, vocoder

1 INTRODUCTION

Time-scale modification (TSM) algorithms have been investigated and improved for the purpose of speech stretching for more than 20 years (1, 2). Among these methods, waveform similarity-based overlap-add (WSOLA) algorithm and phase vocoder are commonly applied in time and frequency domain, respectively. WSOLA algorithm selects the composite signal with the previous frame during each iteration; the most similar frame is combined with the next one of audio data to ensure signal continuity. The core of phase vocoder is the short-time Fourier transform (STFT). The STFT converts a time domain representation of signal into a time-frequency representation, thus allowing modification to the amplitude or phase of specific frequency components before re-synthesis back to the time domain by the inverse STFT. In particular, the time evolution of the resynthesized signal can be changed by means of modifying the time position of STFT frames prior to re-synthesis, so time-scale modification can be achieved (3). Speech stretching techniques have also been applied to speech-to-singing synthesis (4, 5), where rhythm and melody are added by TSM and pitch shift respectively to speech utterances.

However, stretching uniformly on speech may cause unnaturalness since the vowel part could be stretched a lot but the consonant part should not. A non-uniform time-scaling algorithm has been proposed to preserve the quality of the speech by applying vowel detection (6). For Mandarin as well as in other languages, automatic parsing speech into vowel and consonant parts is still an active research area (7, 8). In addition, because of the transition between consonants and vowels, an appropriate stretch region may also include non-vowel region of utterance. In this paper, we propose a non-uniform stretch method to find 'optimal solution' for stretching Mandarin speech, and demonstrate the results by adding rhythm and tempo in the spoken sentences. A high quality speech analysis/synthesis tool called World vocoder is used to extract speech features (9), namely the fundamental frequency (f0), the spectral envelope (SP) and the aperiodic parameter (AP). They are estimated by HARVEST (10), CheapTrick (11) and D4C (12) algorithms, respectively.

*Email: phycause@gapp.nthu.edu.tw

The rest of this paper is organized as follows. In Sec. 2, the proposed algorithm to decide the stretch region is described. Sec. 3, data preparation and recording process are introduced. Sec. 4 includes the result chart of subject listening tests. The pros and cons of the proposed work are summarized in Sec. 5.

2 PROPOSED STRATEGY

The goal of our strategy is to find the optimal stretchable region of utterances and add rhythm to a sentence of spoken speech. In this section, we will give an overview of the method at first, and describe the details in the following paragraphs.

2.1 Overview

An overview of the proposed framework for natural Mandarin spoken word stretching is shown in Figure 1. It consists of dictionary creation and signal modification. In the dictionary creation part, pairs of regular-speed and slow utterances are recorded and analyzed by the World vocoder. Two examples of the features extracted by the vocoder, which are SP and AP, are shown as Figure 2. Next, the regular-speed utterance which is stretched by linear interpolation in the vocoder domain (LI-VD) is compared to the slow-speed utterance in the SP domain. Among all possible choices of the stretch region, we can find the optimal one which has minimum loss between the SP of the stretched regular-speed and the slow-speed utterances, and the optimal dividing point is defined to be the beginning of the optimal stretch region. In the signal modification part, the spoken sentence is segmented into utterances at the beginning. Then, the optimal point of each spoken word in full sentences are aligned with regular-speed utterances having same pronunciations by dynamic time warping (DTW). Furthermore, the optimal region of speech is then stretched by LI-VD in accordance with specific rhythm and tempo. The last step is to synthesize and concatenate the stretched Mandarin speech to form a full sentence.

2.2 Time stretching

Linear interpolation in the vocoder domain (LI-VD) is used to stretch the spoken word. In this method, the speech features are extracted by World vocoder, and the features are interpolated along the time axis linearly. Then, the time domain stretched signals are synthesized by the vocoder.

To describe it more precisely, suppose that a vector-valued time series $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_p)$ is given and the corresponding time of occurrence is $\boldsymbol{t} = (t_1, t_2, ..., t_p)$. We can calculate the new time series feature $\boldsymbol{\gamma'} = (\gamma'_1, \gamma'_2, ..., \gamma'_q))$ based on the new time axis $\boldsymbol{t'} = (t'_1, t'_2, ..., t'_q))$ by the following method: For each t'_i in $\boldsymbol{t'}$, we calculate its corresponding coordinate τ_i in $\boldsymbol{t}$ through the Equation (1). Let t_a and t_b be the two points along the time axis that are closest to τ_i, and then γ'_i can be derived as follows,

$$\tau_i = (t_p - t_1)\frac{t'_i - t'_1}{t'_q - t'_1} + t_1 \tag{1}$$

$$\gamma'_i = (\gamma_b - \gamma_a)\frac{\tau_i - t_a}{t_b - t_a} + \gamma_a. \tag{2}$$

2.3 Creating dictionary

In creating the dictionary, we prepared three features to use in the signal modification part – optimal dividing point, pronunciation and SP of regular-speed utterance of every word. The region between the optimal dividing point and the end of the utterance is the section which could be stretched; and the rest of the utterance, which is the region before the optimal dividing point, remains unchanged.

There are three steps to find the optimal dividing point. First, we prepared pairs of regular-speed and slow-speed utterances with same pronunciation, and then extracted their SP by the World vocoder. Next, we exhaustively test the fitness of setting a frame in regular-speed utterance as the dividing point, by concatenating the region

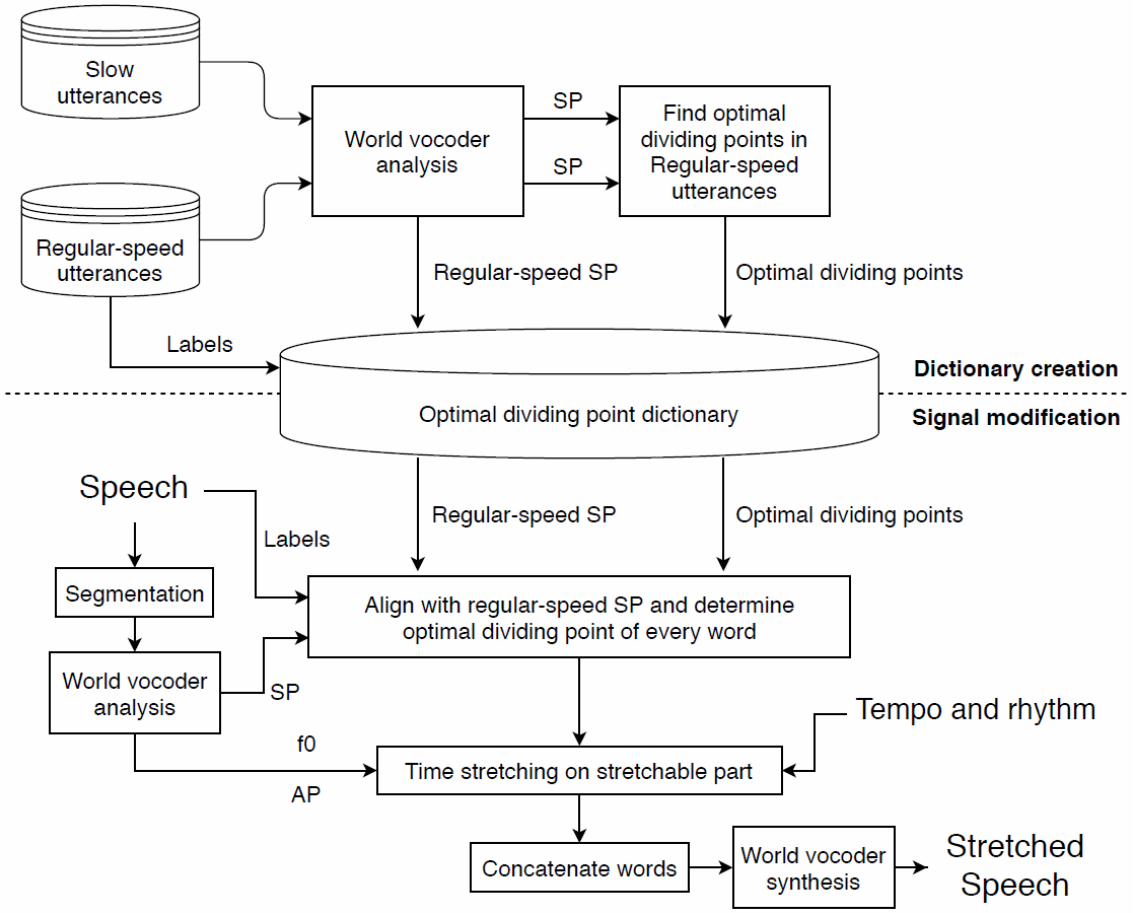


Figure 1. The proposed framework for stretching a sentence of Mandarin speech naturally.

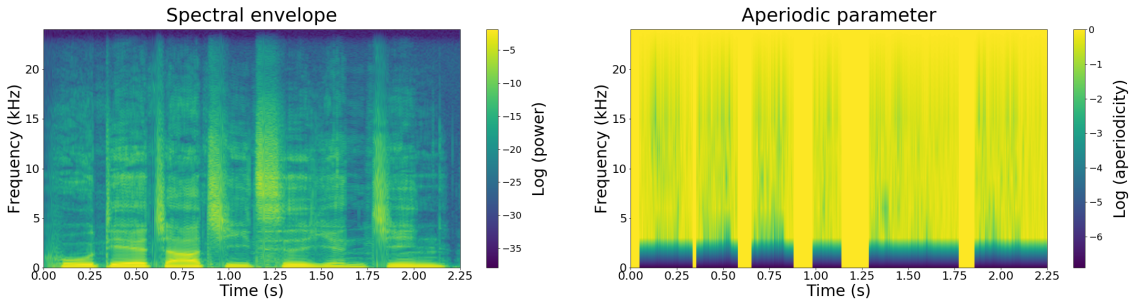
before the frame with original speed while stretching the following region to make the utterance the same length as the slow-speed utterance. The time stretching method (LI-VD) is described in Sec. 2.2.

Last, the Euclidean distance of SP between the stretched regular-speed and slow-speed utterance was calculated, which was defined to be *the loss*. Since the loss could be plotted as a function of the dividing-point location, the dividing point with the minimum loss is regarded as the optimal dividing point of the utterance. Some examples of the loss curve are shown in Figure 3.

2.4 Modification and optimal dividing point alignment

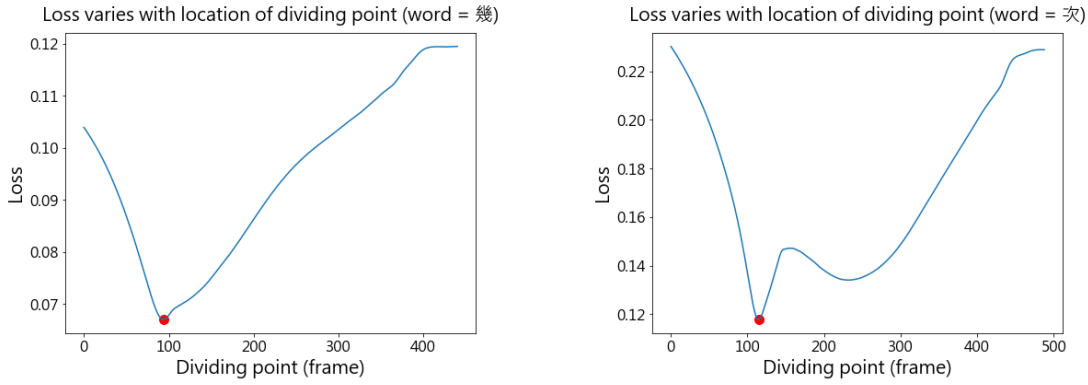
To find the optimal dividing points of each word that occurs in full-sentence fluent speech, the dynamic time warping (DTW) algorithm is used (13, 14). Suppose that two time series $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_L)$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_M)$ are given. Then, an Euclidean distance $d(\alpha_i, \beta_j)$ is defined between any pair of data points, an *accumulative matrix* $D_{i,j}$ is assigned for all $0 \leq i \leq L$ and $0 \leq j \leq M$ as follows.

$$D_{i,j} = \begin{cases} +\infty, & \text{if } i = 0 \text{ or } j = 0 \\ d(\alpha_1, \beta_1) \ , & \text{if } i = j = 1 \\ d(\alpha_i, \beta_j) + \min\{D_{i-1,j} \ , D_{i,j-1} \ , D_{i-1,j-1}\}, & \text{otherwise.} \end{cases} \tag{3}$$



(a) (b)

Figure 2. (a) and (b) are the examples of SP and AP estimated by World vocoder respectively, with 4096 samples per frame and 5ms of step time. The sentence Pinyin of the examples is "hú dié zhǎ jǐ cì yǎn jīng".



(a) (b)

Figure 3. Examples of how the loss function varies with the location of the dividing point, and the optimal dividing points are also marked. The Pinyin of the pronunciations of (a) and (b) are "jǐ" and "cì", respectively.

Every point between two time series could be aligned by the resulting optimal warping path based on accumulative matrix $D_{i,j}$. Next, the optimal dividing point on the spoken word in fluent speech is found by aligning the optimal dividing point on the corresponding regular-speed utterance in the dictionary to the spoken word.

3 EXPERIMENTS

3.1 Subjective listening tests

To compare our proposed method for Mandarin speech stretching with other algorithms, we designed a listening test including three sentences. For each sentence, subjects first hear a Mandarin spoken speech with the rhythm and tempo, which is called the ground-truth (GT) audio. The followings are three stretched audios which were synthesized using different algorithms; one of them is WSOLA, one is LI-VD with uniform stretching, and the last one is LI-VD with optimal stretching. Every subject had to choose one audio that he/she considered to be most similar to GT audio. The presentation order of the results from different algorithms was randomized.

3.2 Data preparation

3.2.1 Recording settings

In the experiments, the data were recorded in a room with low reverberation. In addition, the condenser microphone with a pop-filter was used, and 28 dB gain was applied to amplify the signal. Moreover, the signals were sampled at 96 kHz, and then down-sampled to 48 kHz before analysis.

3.2.2 Utterances

The three sentences for the listening tests are excerpts from the lyrics of popular songs. For each sentence, two versions were recorded. The first one is the spoken sentence, which was used to synthesize stretched speech. The second version is the spoken sentence followed by the rhythm and tempo of the original song as the GT. The Chinese characters with Pinyin of these three sentences are as follows, and the corresponding rhythm and tempo are shown in Figure 4.

1. 蝴蝶眨幾次眼睛才學會飛行 (hú dié zhǎ jǐ cì yǎn jīng cái xué huì fēi xíng)
2. 當一陣風吹來風箏飛向天空 (dāng yī zhèn fēng chuī lái fēng zhēng fēi xiàng tiān kōng)
3. 爲了你而祈禱而祝福而感動 (wéi liǎo nǐ ér qí dǎo ér zhù fú ér gǎn dòng)

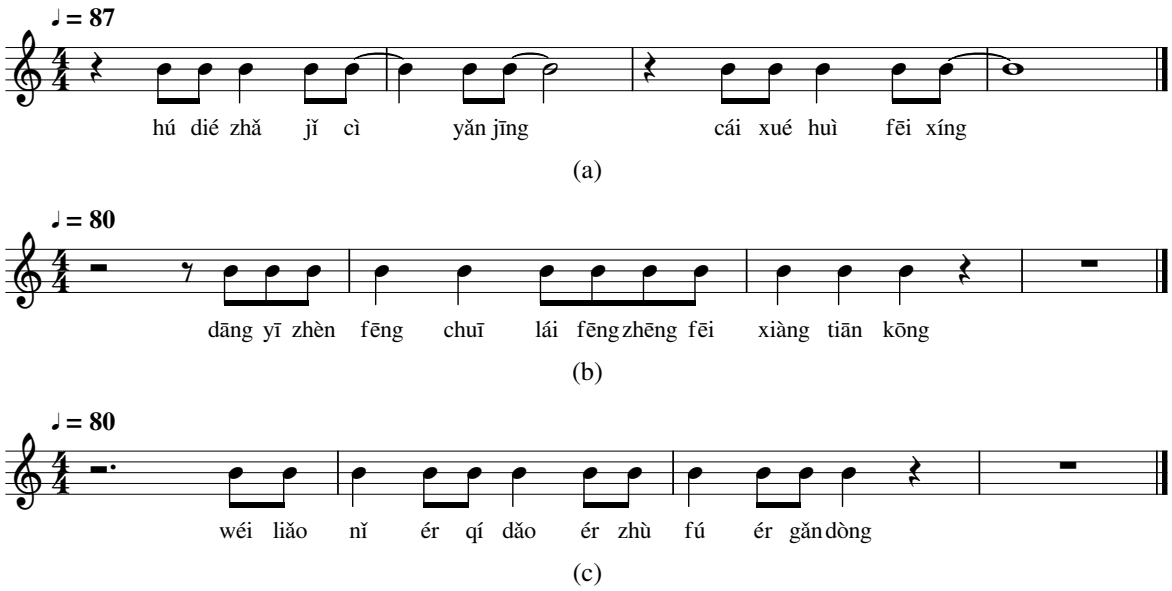


Figure 4. The rhythm and tempo of every sentence in the subjective listening tests. (a), (b) and (c) correspond to sentence 1, 2 and 3, respectively.

In addition, we also recorded the regular-speed and slow-speed utterances of every word that appears in these three sentences to create an optimal dividing points dictionary, which was mentioned in Sec 2.3. Slow utterances were recorded in 3 beats with a tempo at 96 bpm (beats per minute). For each recorded speech, the onset and the end of each word was labeled for the purpose of segmentation.

4 RESULTS

4.1 Subjective listening tests

The results of the listening tests by 41 subjects are summarized in Table 1. In average, 79.6% of answers preferred the speech stretched by the proposed method over the other two methods. This shows that our method outperformed most of the times in terms of providing natural stretching.

Table 1. The results of the listening tests.

Method	Sentence 1	Sentence 2	Sentence 3	Average
Proposed strategy	92.7%	87.8%	58.5%	79.6%
Uniform LI-VD	4.9%	12.2%	31.7%	16.3%
Uniform WSOLA	2.4%	0.0%	9.8%	4.1%

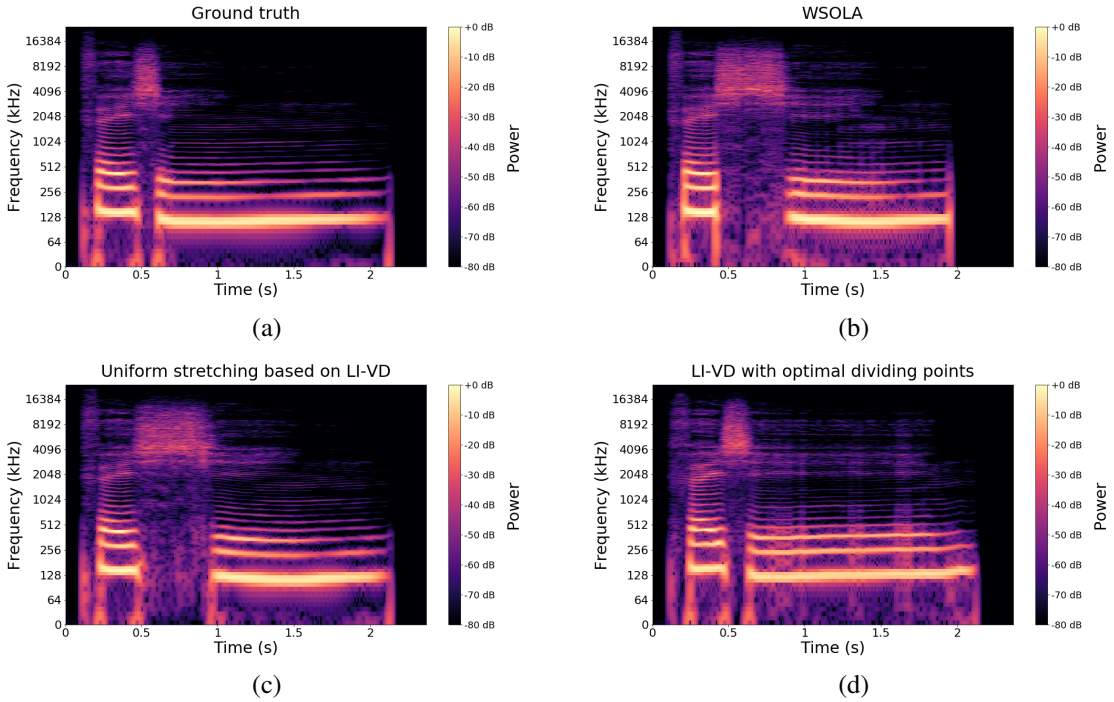


Figure 5. Examples of power spectrograms of the audio files in the listening test. (a) is the ground truth and (b), (c), (d) are the stretched speech with 3 different stretching method. The pronunciation of the examples is "fēi xíng" in Pinyin.

4.2 Discussion

4.2.1 Subjective tests

According to the results, most of the subjects chose the speech stretched by the proposed strategy as the most natural. However, in sentence 3, 31.7% of the subjects preferred the stretched audio using the uniform LI-VD method; in addition, some subjects mentioned that they struggled to differentiate between the speech stretched by uniform and non-uniform LI-VD in sentence 3. By inspecting Figure 4-(c), we suggest that the reason is

because "fú" is the only quarter-note word that starts with a fricative consonant in sentence 3; therefore, the advantage of the proposed stretching strategy is not prominent. In contrast, the other two sentences contain multiple sustained words that start with a fricative or an affricate, such as "cì" and "xíng" in sentence 1, and "fēng", "fēi" in sentence 2, so the subjects could discriminate the most natural synthesized speech from the other ones more easily.

4.2.2 Some observations

Examples of the power spectrogram from listening test audio files are shown in Figure 5. It demonstrates that the consonant regions of the utterances are stretched if the speech is stretched by uniform WSOLA or uniform LI-VD, which are different from the GT. In contrast, the stretched speech with the proposed method looks more similar to the GT. This observation agrees with the results of the listening tests.

5 CONCLUSIONS

A new method is proposed for natural Mandarin spoken utterance stretching. We presented a strategy of finding the optimal dividing point by comparing the regular-speed and slow-speed utterances to define the stretchable region in every utterance, which is the region between the optimal dividing point and the end of the word. Subsequently, the optimal dividing points could be aligned to every word in the full sentence to decide the stretched region. This method is preferred according to the outcome of the subjective listening tests. Having demonstrated the effectiveness of the current methods, in the future, all possible word pronunciations (including the different tones) in Mandarin are going to be recorded so a comprehensive optimal dividing-point dictionary will be built.

ACKNOWLEDGEMENTS

This research is supported by the Ministry of Science and Technology of Taiwan under grant No. 108-2634-F-007-003).

REFERENCES

[1] Driedger, J. Time-scale modication algorithms for music audio signals, Master's thesis, Saarland University, 2011.

[2] Moulines, E.; Laroche, J. Non-parametric techniques for pitch-scale and time-scale modification of speech, Speech Communication, Vol 16 (2), 1995, pp 175–205.

[3] Driedger, J.; Meinard, M. TSM toolbox: Matlab implementations of time-scale modification algorithms, Proc. of the International Conference on Digital Audio (Proc. of DAFx), 2014, pp 249-256.

[4] Saitou, T; et al. Speech-To-Singing Synthesis: Converting Speaking Voices to Singing Voices by Controlling Acoustic Features Unique to Singing Voices, Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (Proc. WASPAA), 2007, pp 215–218.

[5] Aso S.; Saitou T.; Goto M.; Itoyama K.; Takahashi T.; Komatani K.; et al. Speakbysinging: converting singing voices to speaking voices while retaining voice timbre, Proc. of the 13th International Conference on Digital Audio Effects (DAFx-10), 2010.

[6] Kupryjanow, A.; Czyzewski, A. A non-uniform real-time speech time-scale stretching method, Proc. of the International Conference on Signal Processing and Multimedia Applications (SIGMAP), 2011, pp 1-7.

[7] Lin, M.-T.; Lee, C.-K.; Lin, C.-Y. Consonant/vowel segmentation for Mandarin syllable recognition, Comp. Speech and Lang., Vol 13 (3), 1999, pp 207–222.

[8] Liu, Y.-T.; Tsao, Y.; Chang, R.-Y. A deep neural network based approach to Mandarin consonant/vowel separation, IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW), 2015.

[9] Morise, M.; Yokomori, F.; Ozawa, K. WORLD: a vocoder based high-quality speech synthesis system for real-time applications, IEICE Trans. Inf. Syst., Vol E99-D (7), 2016, pp 1877–1884.

[10] Morise, M. Harvest: A high-performance fundamental frequency estimator from speech signals, Proc. Interspeech, 2017, pp 2321–2325.

[11] Morise, M. CheapTrick, a spectral envelope estimator for high-quality speech synthesis, Speech Communication, Vol 67, 2015, pp 1-7.

[12] Morise, M. D4C, a band-aperiodicity estimator for high-quality speech synthesis, Speech Communication, Vol 84, 2016, pp 57–65.

[13] Keogh, E.; Ratanamahatana, A. C. Exact indexing of dynamic time warping. Knowledge and Information Systems, Vol 7 (3), 2005, pp 358–386.

[14] Kruskal, J. B. An overview of sequence comparison: Time warps, string edits, and macromolecules. SIAM Review, Vol 25 (2), 1983, pp 201–237.