JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 36, 547-559 (2020)
DOI: 10.6688/JISE.202005_36(3).0005

Automatic Methods for Detecting Sung Lyrics Error*

WEI-HO TSAI AND SHIANG-SHIUN KUNG
Department of Electronic Engineering
National Taipei University of Technology
Taipei, 106 Taiwan
E-mail: whtsai@ntut.edu.tw; squarprince@yahoo.com.tw

A sung lyrics error detection system is proposed to examine if the lyrics sung by a singer are incorrect, thereby providing a clue for singing skill evaluation. In essence, sung lyrics error detection is similar to the problem of speech utterance verification in the speech recognition research community, and therefore the techniques in the latter can be applied to the former. However, our experiment found that a speech utterance verification system is far from capable of handling singing data, mainly because of the significant difference between singing and speech. To tackle this problem, we develop two strategies, respectively, from a signal processing perspective and from a model processing perspective. In the signal processing perspective, we recognize that the vowels are often lengthened during singing, and thus propose vowel shrinking/decimation to adjust the length of a vowel in singing to a normal length in speaking. In the model processing perspective, we combine a duration modeling concept into the acoustic modeling to reduce the differences between singing and speech. Our experiments show that the proposed methods can improve the performance of the sung lyrics error detection noticeably, compared to a baseline system based on speech utterance verification.

Keywords: duration modeling, singing, speech, sung lyrics, utterance verification

1. INTRODUCTION

Humans have an instinct to sing. However, great singing is a skill that needs to be cultivated. In addition to practice and practice, it is a necessity for singers to know how well or bad they sing, thereby learning how to improve. Usually, singers rely on the comments from other people to know how well or bad they sing. For singing learners, they may even need a tutor to help them sing better. Sometimes learning to sing can be costly and ineffective. To provide singing learners with a handy assistance, machine-based evaluation of singing performance could be a promising solution, since it can be done at anytime and anywhere. However, machine-based singing evaluation is still not available or acceptable at this stage, since the associate techniques are not well investigated.

Most of the current karaoke apparatuses have a function of singing evaluation. However, their evaluation results are far from acceptable or even like a random score for fun. Although there have been several studies [1-14] to this technique, most of them are reported in patent documentation, which only describe their implementation details and fail to present the theoretical foundation and qualitative analysis conducted to validate their methods. Only very few studies are reported in scientific literature. The most thorough investigation of this research topic is a work reported in [8]. It comprehensive-

Received September 6, 2018; revised November 5, 2018; accepted January 2, 2019.
Communicated by Chung-Hsien Wu.
* This work was supported in part by Ministry of Science and Technology, Taiwan, under Grant No. MOST 106-2221-E-027-125-MY2.

ly discusses the strategies and acoustic cues for singing evaluation. However, the methods proposed in [8] focus only on measuring the correctness of a singing performance in terms of pitch, volume, and rhythm, while ignores the lyrics sung by a singer. When a singer sings behind or ahead of the beat, it may arise from or lead to wrong sung lyrics. Thus, measuring the correctness of the sung lyrics is also an indispensable part of singing evaluation. Yet, to the best of our knowledge, no prior study has discussed this issue. Thus, this work attempts to investigate and propose solutions to this issue.

To be specific, the goal of this study is to check if there are errors in the sung lyrics and where the errors are in a given singing recording. The task is similar to the speech utterance verification [15-21] discussed in the community of speech recognition research. However, given certain lyrics, signals resulted from singing the lyrics can be rather more diverse than speaking the lyrics. For example, the signal of singing a word may be five times length of speaking the same word, or may be sometimes one fifth length of speaking the same word, because the signal length of singing a word depends on the song's melody, tempo, *etc.* As a consequence, sung lyrics verification (error detection) is much more challenging than speech utterance verification, since singing can be considered as an irregular distorted version of speech.

On the other hand, there are a few studies on automatic sung lyrics recognition [22-25], which aims to decode a singing signal into phonemes or words. As it is infeasible to acquire a large enough singing database to train phonetic models for singing signals, most of the research in this topic focused on studying how to exploit structure information, *e.g.*, repetition, and composing information in music to help decode singing signals. However, sung lyrics recognition to date remains an extremely challenging task. The phoneme recognition accuracy is quite low, *i.e.*, below 30%, and hence the technique is not ready to be used in real applications. Recognizing the above-mentioned problems, this study proposes several methods to improve a speech utterance verification system to better handle singing signals.

Theoretically, a sung lyrics error detection system trained using singing data would be better than it is trained using speech data. However, due to the huge variety of acoustic characteristics in singing, there are infinite possibilities of singing rendition for every word. Under this circumstance, it is infeasible to collect sufficient data that cover various possibilities of singing rendition to generate singing phone models. Even though there are numerous songs in the world available for training, a vast majority of songs contain background accompaniments in most or all vocal passages, making them difficult to be used here. Recognizing this, we design the sung lyrics error detection system by improving a speech utterance verification system, rather than starting from scratch by using singing data directly. Two strategies, respectively, from a signal processing perspective and from a model processing perspective are developed. In the signal processing perspective, we recognize that the vowels are often lengthened during singing, and thus propose vowel shrinking/decimation to adjust the length of a vowel in singing to a normal length in speaking. In the model processing perspective, we include a duration model concept in the acoustic modeling to reduce the differences between singing and speech.

The remainder of this paper is organized as follows. In Section 2, we describe a baseline sung lyrics error detection system based on a conventional speech utterance verification approach. Section 3 presents an improved system based on the signal pro-

cessing strategy, namely, vowel shrinking or vowel decimation. Section 4 presents another improved system based on the duration modeling. Section 5 discusses our experiment results. Then, in Section 6, we present our conclusions and indicate the direction of our future work.

2. BASELINE SYSTEM: SUNG LYRICS ERROR DETECTION BASED ON SPEECH UTTERANCE VERIFICATION

As mentioned earlier, sung lyrics error detection is similar to the problem of speech utterance verification in the speech recognition research community. We therefore begin this research by using a conventional speech utterance verification system to perform sung lyrics error detection.

For a test singing recording **S**, our task is to examine if the lyrics sung in the recording are the same as the specified lyrics. The task can be formulated as a statistical hypothesis testing problem, involving null hypothesis H_0 against the alternative hypothesis H_1 as

H_0: the sung lyrics are truly the specified lyrics,
H_1: the sung lyrics are not the specified lyrics.

The optimum test to decide between these two hypotheses is a likelihood ratio test given by

$$\frac{\Pr(\mathbf{S}\,|\,H_0)}{\Pr(\mathbf{S}\,|\,H_1)} \begin{array}{l} \text{accept } H_0 \\ > \\ \leq \\ \text{accept } H_1 \end{array} \delta, \tag{1}$$

where δ is a tunable threshold, $\Pr(\mathbf{S}|H_0)$ is the probability, also referred to as the likelihood for the hypothesis H_0 evaluated for singing recording **S**. The likelihood function for H_1 is likewise $\Pr(\mathbf{S}|H_1)$. To represent H_0 and H_1 mathematically, the sounds belonging to and not belonging to the specified lyrics are respectively modeled by a set of parameters.

Our system is based on context-dependent Hidden Markov Model (HMM) sets with Gaussian mixture output distributions. Parameters of an HMM consist of initial state probabilities, state transition probabilities, mixture weights, mean vectors, and covariance matrices. The system is built using the HMM Toolkit (HTK) [26]. Prior to modeling, the acoustic data is represented by a stream of 39 dimensional feature vectors with a frame spacing of 10ms, using `MFCC_E_D_A_Z` defined in HTK. Briefly, `MFCC_E_D_A_Z`, which consists of zero-meaned Mel-Frequency Cepstral Coefficients (MFCCs) appended with delta coefficients, acceleration coefficients, and log energy, is computed. In this study, we only consider Mandarin speech and singing data, but the system could be extended to handle other languages. The training data used for our speech recognition system stem from TCC-300 [27], which is composed of Mandarin speech utterances recorded in quiet environments.

There are 151 context-dependent sub-syllable phones used in this study. Each individual phone is represented by an HMM. In addition to the 151 phones, the acoustic

model set contains two silence models, one for silence, and one for short inter-word pauses with the latter preserving context across words. On the other hand, for the sake of training efficacy, we use the state-tying technique provided by HTK for each phone with different context.

As shown in Fig. 1, when a test singing recording is input, the system computes its feature vectors and concatenates several phone models to form a phone sequence model Λ, according to the specified lyrics and a pronunciation dictionary. The phone sequence model is then used to evaluate the log likelihood ratio for the feature vectors:

$$\mathrm{R}(\mathbf{O}) = \ln \Pr(\mathbf{O} \mid \Lambda) - \ln \Pr(\mathbf{O} \mid \Lambda^*) \begin{array}{c} \text{correct} \\ \gtrless \\ \text{incorrect} \end{array} \delta, \tag{2}$$

where $\mathbf{O}$ is the 39 dimensional feature vectors extracted from the test singing signal, Λ^* represents the most likely phone sequence model obtained with the free syllable decoding based on Viterbi algorithm, and δ is a tunable threshold. Thus, the test sung lyrics are hypothesized to be incorrect, if the ratio R($\mathbf{O}$) is smaller than the pre-set threshold δ.

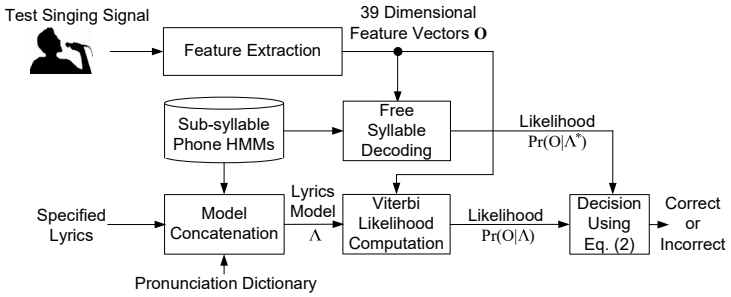


Fig. 1. A baseline sung lyrics error detection system, based on the conventional speech utterance verification approach.

3. IMPROVED SYSTEM BASED ON THE SIGNAL PROCESSING STRATEGY

Our experiment results, detailed in Section 5, show that a speech utterance verification system cannot handle singing data well, mainly because of the significant characteristic differences between singing and speech. One major difference, which deteriorates the performance of a speech utterance verification system severely, is the lengthening[+] of vowels in singing. Fig. 2 shows a speech signal and a singing signal produced by a person who read and sang the same lyrics, respectively. We can see from Fig. 2 that the lengths of vowels are noticeably different between speech and singing, whereas the lengths of consonants are roughly similar between speech and singing. Hence, to solve the problem of vowel lengthening, we propose two approaches to adjust the length of a vowel in singing to a normal length in speaking, namely, vowel shrinking and vowel decimation.

[+] It is found that the shortening of vowels in singing does not have significant impact on the detection, and thus we only deal with the problem of vowel lengthening in singing.

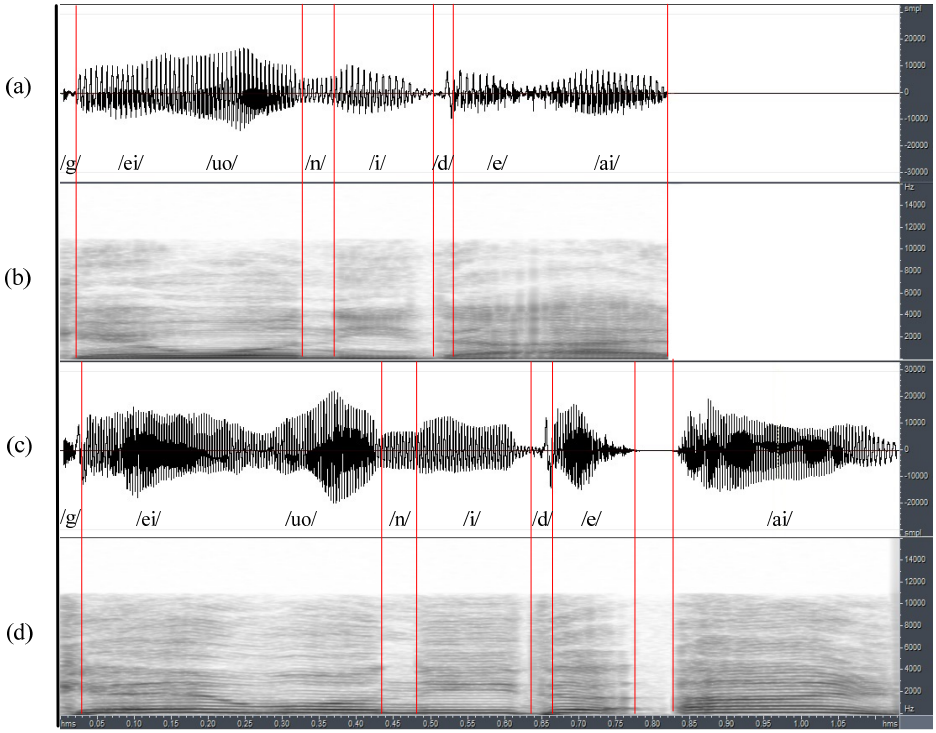


Fig. 2. A speech signal and a singing signal produced by a person who read and sang lyrics "給我你的愛", respectively, where (a) is the speech waveform; (b) is the speech spectrogram; (c) is the singing waveform; and (d) is the singing spectrogram.

The method begins by finding the intervals where vowels occur in a singing signal. Since vowel signals are periodic, we can compute the pitch of each short segment of the test singing signal and check the periodicity by evaluating the pitch value. If the pitch value is larger than a threshold, then the segment is hypothesized as a vowel; otherwise, the segment is hypothesized as a consonant. In our implementation, we use Yet Another Algorithm for Pitch Tracking (YAAPT) [28] to compute the pitch of singing signal.

As shown in Fig. 3, once an interval of vowel (consisting of several vowel segments) is found, the system evaluates its duration. If the length is larger than a threshold d frames,

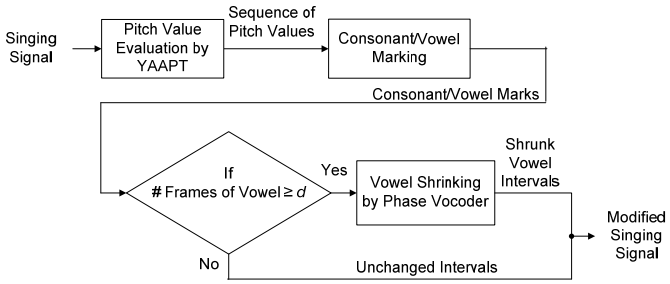
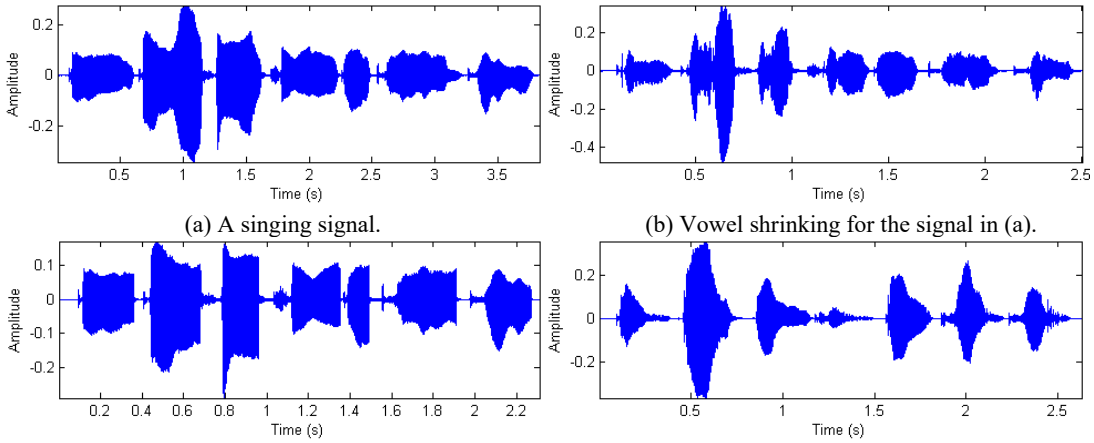


Fig. 3. Vowel shrinking by phase vocoder.

the system adjusts its length by a factor of k. One of the adjusting approaches, vowel shrinking, is based on phase vocoder [29], which performs time-scaling and pitch-shifting on the signal. For a vowel having length of T, the Phase Vocoder shrinks its length to T/r so that the vowel sounds like speeding up. Another approach, vowel decimation, is to decimate the repetitions in vowels. More specifically, the system cuts out the backend of a vowel and only preserves the frontend T/r-length part of the vowel.

Fig. 4 shows an example of a singing signal before and after vowel length adjusting. It can be seen from Fig. 4 that the length of a singing signal after vowel shrinking or vowel decimation has been adjusted close to that of the speech signal generated with the lyrics same as the singing signal. In the experiment section, we evaluate the two approaches for the sung lyrics error detection problem, and refer to them as vowel shrinking approach and vowel decimation approach, respectively. The improved sung lyrics error detection system based on the vowel shrinking/decimation is shown in Fig. 5, where the vowel shrinking/decimation is performed prior to the feature extraction.



(a) A singing signal. (b) Vowel shrinking for the signal in (a).

(c) Vowel decimation for the signal in (a). (d) Speech signal based on the lyrics the same as that in (a).

Fig. 4. An example of a singing signal before and after vowel shrinking or vowel decimation and its counterpart speech signal based on the same lyrics.

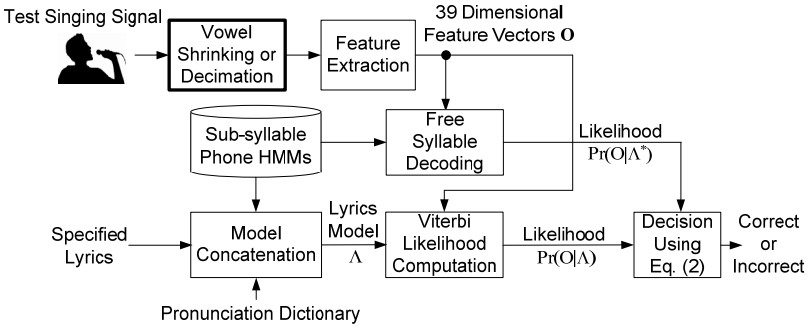


Fig. 5. An improved sung lyrics error detection system based on the signal processing approach, *i.e.*, vowel shrinking or vowel decimation.

4. IMPROVED SYSTEM BASED ON THE MODEL PROCESSING STRATEGY

In a conventional speech utterance verification system, a consonant phone model contains two to three states, and a vowel phone model contains two to five states in general. The states are with left-to-right topology to capture the sequential information of an acoustic signal. However, when Viterbi algorithm is performed, a singing vowel signal may have too many frames to be assigned to the states within a vowel phone model, since the probability that multiple consecutive frames staying in the same state is exponentially decaying with the number of frames. In other words, two to five states in a vowel phone model may not be able to cover the great number of frames in a singing vowel. To solve this problem, we propose using the concept of duration modeling [30] in the testing phase.

As shown in Fig. 6, we duplicate each state as K states to absorb the great number of frames likely in a singing vowel, where K is set to 4 empirically in our experiment. It is found that the duplication of state is not detrimental for the case when singing is faster than the normal speech, since the length of a vowel is more than 4 frames. The improved sung lyrics error detection system based on the duration modeling is shown in Fig. 7.

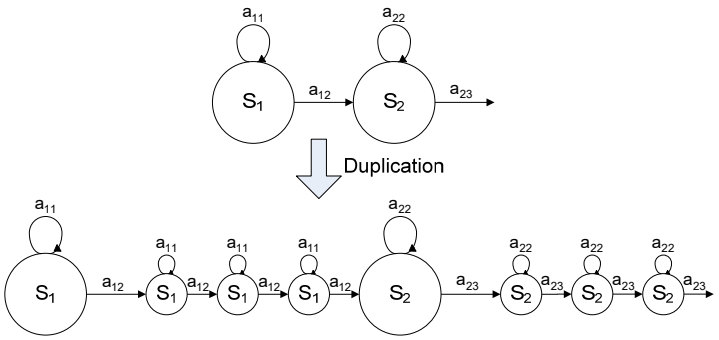


Fig. 6. Concept of the duration modeling by duplicating each state as four states, in which a_{ij} represents the transition probability from state S_i to S_j, $i, j = 1$ and 2 in this case.

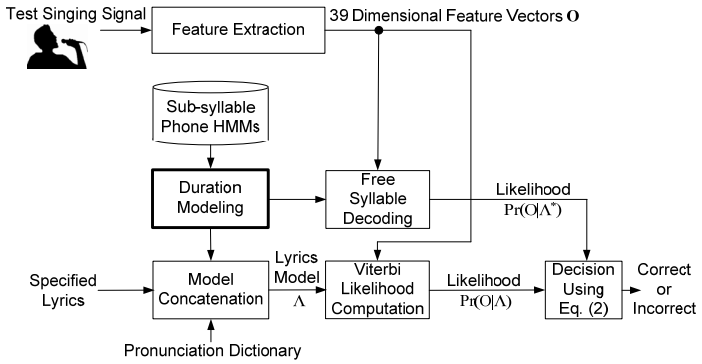


Fig. 7. An improved sung lyrics error detection system based on the duration modeling.

5. EXPERIMENTS

5.1 Databases

The audio data used in this study involved two databases, one is TCC-300 and the other is collected by ourselves, denoted by DB-S. The TCC-300 was used to establish the speech recognizer described in Section 2, and DB-S was used to test our sung lyrics error detection system.

In collecting DB-S, we invited six participants, including one male and five female participants, between the ages of 20 and 29 to produce vocal recordings. First, we asked each participant to sing 15 passages of Mandarin pop songs using a Karaoke machine in a quiet room. There are 7 slow songs and 8 fast songs. All the passages were recorded at 22.05 kHz, 16 bits, in mono PCM wave. The Karaoke accompaniments were output to a headset and were not captured in the recordings. The duration of each passage ranges from 17 to 26 seconds. The recordings were denoted by DB-S-1.

Then, we simulated the situations that some mistakes occur in the sung lyrics. As shown in Table 1, there were four situations considered in our study, namely, partial lyrics are wrong, partial lyrics are in a reverse order, partial lyrics are missing, and wrong repetition is appended. The resulting recordings in the four situations were denoted by DB-S-2, DB-S-3, DB-S-4, DB-S-5, respectively. The rules for generating the four types of mistakes are also described in Table 1.

Next, we segmented each recording (passage) into small phrases, each in range from 5 to 13 second. The phrases were used as the testing samples to evaluate the performance of the sung lyrics error detection system. There were 600 phrases for each subset.

Table 1. Simulated mistakes in sung lyrics.

Subset	Situation	Rules for Simulating Singing Mistakes	Examples of the Simulated Mistakes for Lyrics in Chinese Characters: “等到風景都看透”
DB-S-2	Partial lyrics are wrong	One of the nouns or verb subphrases in a phrase is replaced by an arbitrary noun or verb subphrase.	“等到人生都看透”
DB-S-3	Partial lyrics are in a reverse order	Two subphrases are interchanged	“都看透等到風景”
DB-S-4	Partial lyrics are missing	One of the nouns or verb subphrases in a phrase is taken out.	“等到風景都”
DB-S-5	Wrong repetition is appended	One of the nouns or verb subphrases in a phrase is repeated.	“等到風景都看透看透”

5.2 Experiment Results

There may be two types of error in the results of sung lyrics error detection. One is

called False Alarm (FA), which means that a singing error detected by the system does not exist in the test recording. The other one is called Miss Detection (MD), which means that a singing error occurring in the test recording is not detected by the system. The two types of error are trade-off and can be presented with better visualization by the DET plot [31]. Fig. 8 shows the DET curves obtained with the baseline system, *i.e.*, sung lyrics error detection based on the conventional speech utterance verification approach described in Section 2.

In Fig. 8, the equal error rate (EER), which is the probability at which the percentage of FA is equal to the percentage of MD, is 22.5%, 19.5%, 47.0%, and 39.7%, respectively, for the four situations: (a) partial lyrics are wrong; (b) partial lyrics are in a reverse order; (c) partial lyrics are missing; and (d) wrong repetition is appended. We can see from Fig. 8 that the mistake situations which is the most difficult to detect is (c) partial lyrics are missing. Overall, it is clear the detection results are far from acceptable, especially for situations (c) and (d), as the detection performance is just the level of random guess.

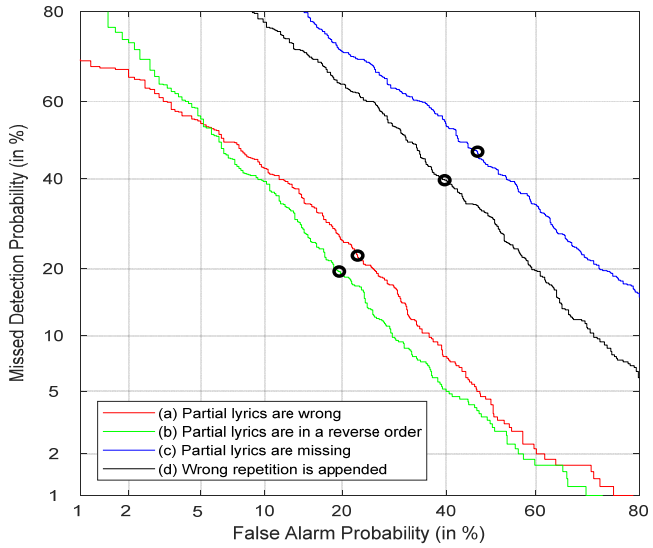
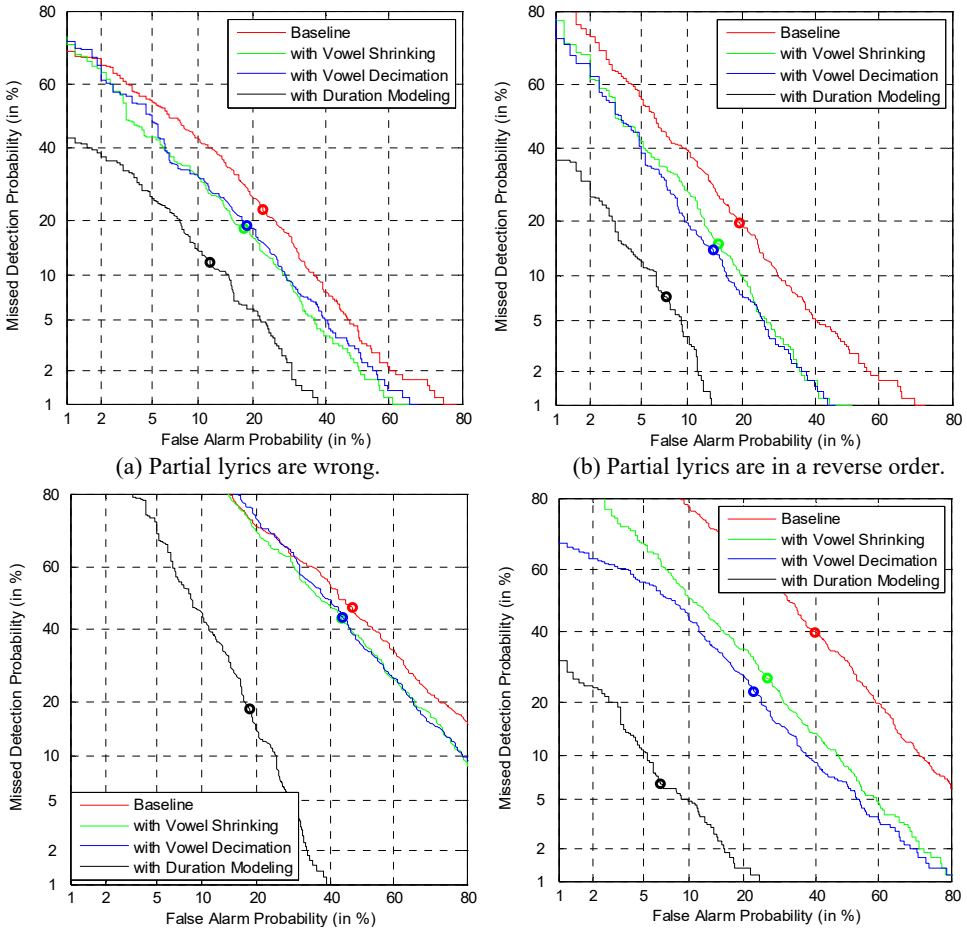


Fig. 8. DET curves obtained with the baseline system, *i.e.*, sung lyrics error detection based on the conventional speech utterance verification approach described in Section 2. The circle on each curve indicates the equal error rate (EER).

Next, we evaluated the performances of the proposed sung lyrics detection methods for improving the baseline system. Fig. 9 shows the detection results. The values of d and r used in the vowel shrinking/decimation were set to 80 and 3, empirically. We can see from Fig. 9 that all the proposed methods, namely, with vowel shrinking, with vowel decimation, and duration modeling, does improve the detection performance. The system with vowel decimation performs roughly equal to or even better than the system with vowel shrinking, though the former is much simpler than the latter. In addition, the system with duration modeling is superior to the other three systems. Table 2 summarizes the EERs achieved with the four systems. Compared to the results of the base-

line system, the duration modeling improves the sung lyrics error detection performance significantly, with more than 45% error reduction for all the mistake situations. Fig. 10 shows the overall performance of the sung lyrics error detection with the duration modeling method for all the situations. The EER is 11.3%. This confirms the feasibility of examining sung lyrics with our improved speech utterance verification framework.



(a) Partial lyrics are wrong.

(b) Partial lyrics are in a reverse order.

(c) Partial lyrics are missing.

(d) Wrong repetition is appended.

Fig. 9. DET curves obtained with the baseline system and the proposed systems.

Table 2. EER (%) obtained with the baseline system and the proposed systems.

Condition	Baseline System	With Vowel Shrinking	With Vowel Decimation	With Duration Modeling
Partial lyrics are wrong	22.5	18.2	18.8	11.8
Partial lyrics are in a reverse order	19.5	15.2	14.2	7.3
Partial lyrics are missing	47.0	43.5	43.8	18.5
Wrong repetition is appended	39.7	26.0	22.5	6.5

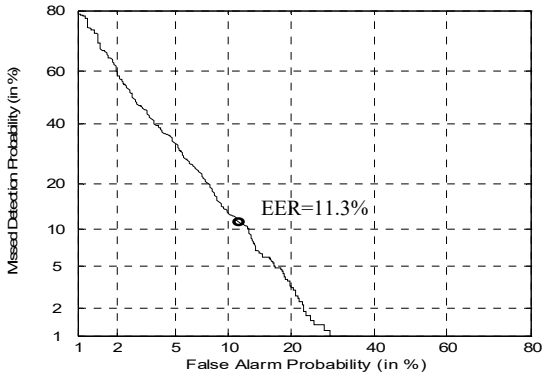


Fig. 10. Overall performance of the sung lyrics error detection with the duration modeling method for all the mistake situations.

6. CONCLUSIONS

In this study, a sung lyrics error detection system has been developed to examine if the lyrics sung by a singer are correct or not. Although sung lyrics error detection is similar to the problem of speech utterance verification, our experiment found that a speech utterance verification system is far from capable of handling singing data, mainly because of the significant differences between singing and speech. Thus, we propose two improved strategies. One is to perform vowel shrinking/decimation to adjust the length of a vowel in singing to a normal length in speaking. The other one is to combine a duration model concept into the acoustic modeling to reduce the differences between singing and speech. Our experiment shows that the proposed methods can improve the performance of the sung lyrics error detection in a great level.

The result in such a pilot investigation is encouraging and lays a good foundation for the future development of a singing skill evaluation system. To maximize its practicability and applicability, the first necessity is to further reduce the detection errors. On the other hand, the current system does not consider the background sounds that may exist during singing. When the ambient noise, background vocal, or accompaniments are recorded together with singing, it will be rather challenging for the sung lyrics error detection problem.

REFERENCES

1. T. Tanaka, "Karaoke scoring apparatus analyzing singing voice relative to melody data," US Patent No. 5,889,224, 1999.
2. O. Mayor, J. Bonada, and A. Loscos, "Performance analysis and scoring of the singing voice," in *Proceedings of the 35th International Conference of the Audio Engineering Society*, 2009, pp. 1-7.
3. W. H. Tsai, C. H. Ma, and Y. P. Hsu, "Automatic singing performance evaluation using accompanied vocals as reference bases," *Journal of Information Science and Engineering*, Vol. 31, 2015, pp. 821-838.
4. J. G. Hong and U. J. Kim, "Performance evaluator for use in a karaoke apparatus," US Patent No. 5,557,056, 1996.

5. K. S. Park, "Performance evaluation method for use in a karaoke apparatus," US Patent No. 5,715,179, 1998.
6. H. M. Wang, "Scoring device and method for a karaoke system," US Patent No. 6,326,536, 2001.
7. T. Nakano, M. Goto, and Y. Hiraga, "MiruSinger: a singing skill visualization interface using real-time feedback and music CD recordings as referential data," in *Proceedings of IEEE International Symposium on Multimedia*, 2007, pp. 75-76.
8. W. H. Tsai and H. C. Lee, "Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features," *IEEE Transactions on Audio*, *Speech*, *and Language Processing*, Vol. 20, 2012, pp. 1233-1243.
9. C. S. Park, "Karaoke system capable of scoring singing of a singer on accompaniment thereof," US Patent No. 5,567,162, 1996.
10. B. Pawate, "Method and system for karaoke scoring," US Patent No. 5,719,344, 1998.
11. T. Nakano, M. Goto, and Y. Hiraga, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," in *Proceedings of International Conference on Spoken Language Processing*, 2006, pp. 1706-1709.
12. P. Lal, "A comparison of singing evaluation algorithms," in *Proceedings of International Conference on Spoken Language Processing*, 2006.
13. P. C. Chang, "Method and apparatus for karaoke scoring," US Patent No. 7,304,229, 2007.
14. C. Gupta, H. Li, and Y. Wang, "Perceptual evaluation of singing quality," in *Proceedings of APSIPA Annual Summit and Conference*, 2017, pp. 577-586.
15. H. Jiang and C. H. Lee, "A new approach to utterance verification based on neighborhood information in model space," *IEEE Transactions on Speech and Audio Processing*, Vol. 11, 2003, pp. 425-434.
16. A. Sankar and S. L. Wu, "Utterance verification based on statistics of phone-level confidence scores," in *Proceedings of IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*, 2003, pp. I-584-587.
17. M. H. Siu, B. Mak, and W. H. Au, "Minimization of utterance verification error rate as a constrained optimization problem," *IEEE Signal Processing Letters*, Vol. 13, 2006, pp. 760-763.
18. I. Lane, T. Kawahara, T. Matsui, and S. Nakamura, "Out-of-domain utterance detection using classification confidences of multiple topics," *IEEE Transactions on Audio*, *Speech*, *and Language Processing*, Vol. 15, 2007, pp. 150-161.
19. J. S. Park, G. J. Jang, and J. H. Kim, "Multistage utterance verification for keyword recognition-based online spoken content retrieval," *IEEE Transactions on Consumer Electronics*, Vol. 58, 2012, pp. 1000-1005.
20. T. Kinnunen, M. Sahidullah, I. Kukanov, H. Delgado, M. Todisco, A. Sarkar, N. B. Thomsen, V. Hautamaki, N. Evans, and Z. H. Tan, "Utterance verification for text-dependent speaker recognition," in *Proceedings of Annual Conference of International Speech Communication Association*, 2016.
21. W. Y. Choi, H. J. Song, H. C. J. Kang, and J. G. Park, "I-vector based utterance veryfication for large-vocabulary speech recognition system," in *Proceedings of IEEE International Conference on Computer Communication and the Internet*, 2016, pp. 316-319.

22. T. Hosoya, M. Suzuki, and S. Makino, "Lyrics recognition from a singing voice based on finite state automaton for music information retrieval," in *Proceedings of International Society for Music Information Retrieval*, 2005, pp. 532-535.
23. A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP Journal on Audio*, *Speech*, *and Music Processing*, 2010.
24. A. Mesaros, "Singing voice identification and lyrics transcription for music information retrieval," in *Proceedings of the 7th Conference on Speech Technology and Hu man – Computer Dialogue*, 2013.
25. M. McVicar, D. Ellis, and M. Goto, "Leveraging repetition for improved automatic lyric transcription of popular music," in *Proceedings of IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2014, pp. 3141-3145.
26. The Hidden Markov Model Toolkit, http://htk.eng.cam.ac.uk/.
27. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), http://www.aclclp.org.tw/use_mat_c.php.
28. S. A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking," *The Journal of the Acoustical Society of America*, Vol. 123, 2008, pp. 4559-4571.
29. J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell System Technical Journal*, Vol. 45, 1966, pp. 1493-1509.
30. J. Pylkkönen and M. Kurimo, "Duration modeling techniques for continuous speech recognition," in *Proceedings of IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2004, pp. 385-388.
31. A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings of Eurospeech*, 1997, pp. 1895-1898.

Wei-Ho Tsai (蔡偉和) received his M.S. and Ph.D. degrees in Communication Engineering from National Chiao-Tung University, Hsinchu, Taiwan, in 1997 and 2001, respectively. From 2001 to 2003, he was with Philips Research East Asia, Taipei, Taiwan, where he worked on speech processing problems in embedded systems. From 2003 to 2005, he served as a Postdoctoral Fellow at the Institute of Information Science, Academia Sinica, Taipei, Taiwan. He is currently a Professor in the Department of Electronic Engineering, National Taipei University of Technology, Taiwan. His research interests include spoken language processing and music information retrieval.

Shiang-Shiun Kung (孔祥勳) received the B.S. and M.S. degrees in Electronic Engineering from National Taipei University of Technology, Taipei, Taiwan, in 2014 and 2016, respectively. His research interests include signal processing and multimedia applications.