Interspeech 2018
2-6 September 2018, Hyderabad

Automatic Pronunciation Evaluation of Singing

Chitralekha Gupta[1], *Haizhou Li*[2], *Ye Wang*[1]

[1]School of Computing, and NUS Graduate School for Integrated Science and Engineering
[2]Department of Electrical and Computer Engineering
National University of Singapore

chitralekha@u.nus.edu, haizhou.li@nus.edu.sg, wangye@comp.nus.edu.sg

Abstract

In this work, we develop a strategy to automatically evaluate pronunciation of singing. We apply singing-adapted automatic speech recognizer (ASR) in a two-stage approach for evaluating pronunciation of singing. First, we force-align the lyrics with the sung utterances to obtain the word boundaries. We improve the word boundaries by a novel lexical modification technique. Second, we investigate the performance of the phonetic posteriorgram (PPG) based template independent and dependent methods for scoring the aligned words. To validate the evaluation scheme, we obtain reliable human pronunciation evaluation scores using a crowd-sourcing platform. We show that the automatic evaluation scheme offers quality scores that are close to human judgments.

Index Terms: pronunciation evaluation, lexicon, singing, crowd sourcing

1. Introduction

Automatic pronunciation evaluation of singing is an essential technology in a wide-range of applications. First, lyrics play an important role in music, serving as a cue for detecting a song's identity, or its mood or genre [1, 2]. Therefore, correctly pronouncing the lyrics of a song becomes an important component of a singing performance. In addition, singing is shown to be helpful in improving pronunciation in foreign language learning classes [3, 4]. Evidence from experimental psychology suggests that learning a new language through singing helps improve vocabulary gain, memory recall, and pronunciation [4]. Furthermore, music and speech therapists apply a therapeutic process called Melodic Intonation Therapy (MIT) to treat patients with speech disorders, such as non-fluent aphasia [5].

Computer-aided pronunciation training (CAPT) for speech has been an active area of research [6, 7]. But automatic pronunciation evaluation of singing is still a relatively unexplored area. The state-of-the-art ASR technology cannot be directly applied for singing pronunciation evaluation because of the mismatch between speech and singing. The acoustic characteristics of singing and speech differ in many ways, such as pitch range, vibrato, and phoneme durations [8, 9]. Thus to build a pronunciation evaluation algorithm for singing, the ASR needs to be adapted to singing voice. Moreover, the applicability of the traditional speech pronunciation scoring methods for evaluating singing pronunciation needs to be investigated.

In this work, first, we propose a novel singing-specific lexicon modification method to overcome the vowel duration differences between singing and speech. We hypothesize that this lexicon modification method for obtaining singing-adapted models leads to better word boundary alignment, which is necessary for a reliable scoring. Next, we investigate methods for scoring pronunciation of sung-utterances at word- and song-levels. We believe that incorporating singing-specific characteristics in scoring would yield improved results. Finally, we validate our scores with human judgments. We also verify that crowd-sourcing platforms can be used to obtain reliable human scores for singing evaluation. We report the encouraging experimental results.

2. Related Work

Phonetic errors in non-native (L2) speech are attributed to the influence of the native language (L1) that results in phone substitutions, deletions or insertions [10]. L1 influence also results in phonetic errors in singing of non-native speakers, as reported in [11]. In karaoke-singing, incorrectly pronounced words often occur due to unfamiliar lyrics or song, that results in substitution, deletion, and insertion of words. In this study, we focus on detecting word pronunciation errors which may be due to the influence of L1 or unfamiliarity of the lyrics of the song.

Only a few studies have addressed the problem of evaluating pronunciation of singing. Jha et al. [12] attempted to develop a method for evaluating pronunciation of singing based on vowels. They compared MFCC and pitch-based features to classify sung vowels, and found that there is no significant difference in performance between the two feature sets. However, their work involved manually extracting vowel segments, and also did not extend to consonants. Recently, we studied the difference in pronunciation between speech and singing in South-East Asian English accents, and found that in singing vocals, the consonant errors are more prominent than the vowel errors [11]. We also incorporated the common pronunciation error patterns for a given L1-L2 pair in a dictionary to automatically detect the mispronounced words. But this work is limited by the need of developing an L1-L2 pair specific dictionary, hence cannot be easily generalized.

In traditional CAPT systems, an L1-independent method of scoring a phoneme is the Goodness of Pronunciation (GOP) score, which is the difference between the log-likelihood score from forced alignment and that from open phone loop decoding, where the phone boundaries are obtained from forced-alignment [13, 14]. This is a template (or reference) independent method of scoring. Another method of scoring is template dependent [15, 16, 17], where deep neural net (DNN) phone posteriorgrams (PPG) are used in dynamic time warping (DTW) between a reference utterance and a test utterance to detect word-level mispronunciations. In this work, we investigate how such methods work for singing pronunciation evaluation.

Recently, a large corpus of solo-singing karaoke data called Digital Archive of Mobile Performances (DAMP)[18] was made available for research purposes. However, annotating such data for qualitative tasks such as singing quality assessment or pronunciation quality evaluation, is still a challenging task. We note that crowd-sourcing platforms have been used for labor-intensive tasks such as speech transcription [19], speech

10.21437/Interspeech.2018-1267

quality assessment tasks [20, 21, 22], and speech pronunciation quality assessments [23]. Researchers have found methods to overcome the noisy nature of the data from such platforms, using gold standard questions, and trapping questions [21]. Encouraged by the findings, here we would like to study how to obtain reliable human judgments of singing pronunciation from the crowd-sourcing platform. We validate the crowd-sourced data against a laboratory-controlled listening experiment data.

With the scarcity of large-scale lyrics-aligned singing data, acoustic models for singing pronunciation evaluation can be built by adapting the speech phonetic models to singing. Adaptation of speech models to singing was previously attempted by Mesaros and Virtanen [24, 25] who used the speaker adaptation techniques to transform speech recognizer to singing voice. Similarly, in our previous work [26], we used fMLLR (feature-space maximum likelihood linear regression) and lyrics-aligned transcriptions of a subset of the DAMP dataset for semi-supervised speaker adaptive training (SAT) of the speech models to singing. In this work, we investigate the performance of word-alignment with our proposed duration-based lexicon modification method along with these singing-adapted models.

3. Singing Pronunciation Evaluation

Speech and singing have many similarities because they share the underlying physiological mechanisms for production. This involves similar articulatory movements to produce words in speech and lyrics in singing [27, 28], thus resulting in similar spectral characteristics for the place and manner of articulation of phonemes. Therefore, we adopt the speech pronunciation evaluation methodology for evaluating pronunciation in singing. We evaluate singing pronunciation in a two-stage approach: word alignment, and scoring.

3.1. Word Alignment

Evaluation of pronunciation is based on phonetic segments, therefore accuracy of alignment is important as it is going to affect the scoring accuracy. Force-aligning the lyrical words to singing with a speech acoustic model does not provide good alignment due to the mismatch between speech and singing signals. One main difference between singing and speech is the duration of the vowels. In singing, the vowels are stretched in time to sustain the musical notes, which is dictated by the score. Previously, musical score-informed duration modeling of vowels has been used in speech-to-singing voice conversion [29], singing-to-speech conversion [30], and singing syllable segmentation [31]. Rong Gong et al. [32] have incorporated pitch and vowel spectral distribution templates to align audio to the musical score. However in karaoke singing, many amateur singers may not able to follow the musical scores correctly. Therefore a score-informed method of lyrics-to-audio alignment will not be accurate.

In this work, we incorporate a novel singing-specific lexicon modification strategy to improve the forced-alignment word boundaries in singing.

3.1.1. Lexicon Modification

The vowels in singing could be longer in duration than spoken vowels, because they are dictated by the melodic and rhythmic attributes of the song. Longer duration of vowels can be viewed as a type of pronunciation variation. One method of evaluating pronunciation in speech, called the *extended recognition network (ERN)* [33, 10], enhances the lexicon with the possible and expected pronunciation error patterns in the specific L1-L2 pair, such that the ASR selects the closest matching variant at the time of forced-alignment. In this work, we propose to modify the lexicon to model the duration dynamics of vowels in singing. We modify the lexicon for singing such that there are multiple pronunciation variants of every word that represent different vowel durations. We adopt the strategy of optional repetition (up to 4 times, set empirically) of the vowels so as to allow longer duration of the vowels. For example, the word *sleep* will have the following lexicon variants: [S L IY IY IY IY P], [S L IY IY IY P], [S L IY IY P], [S L IY P]. Such variants are created with respect to every vowel in the word. We expect that this method will result in improvement in force-aligned boundaries and thus the pronunciation scores.

3.2. Scoring

We evaluate how close an uttered segment is to an expected phone, while considering the differences between speech and singing phonemes. We define the template independent [13, 14] and dependent [15, 16, 17] scores, called Pronunciation Evaluation Metric (PEM) scores, based on the Phonetic Posteriorgram (PPG). PPG contains the normalized posterior probability of every phone per frame, obtained from decoding a sung utterance with the singing-adapted acoustic models.

3.2.1. Template Independent

Template independent PEM score (PEM_{ind}) indicates how close the pronunciation of a test sung utterance is to the target lyrics, similar to the GOP scores of the CAPT systems [13, 14]. PEM_{ind} is defined as the ratio of the probability of the target phoneme to the sum of probabilities of the rest of the phonemes, averaged over all the frames within the phoneme:

$$PEM_{ind} = \frac{1}{N}\sum_{i=1}^{N}\frac{P_i(T_p)}{\sum_{\forall k \neq p} P_i(T_k)} \quad (1)$$

where $P_i(T_p)$ is posterior probability of the target phone T_p from PPG for a frame i. And N is the number of frames within the phone boundaries obtained from forced-alignment with the target transcription. A high PEM_{ind} score means the uttered phone is close to the target phone.

3.2.2. Template Dependent

Template dependent PEM score (PEM_{dep}) indicates how close the pronunciation of the test sung utterance is to a reference sung utterance [15, 16, 17]. It is computed as the dot product of the reference PPG vector P_r and the test PPG vector P_t.

$$PEM_{dep} = -\log(P_r.P_t) \quad (2)$$

If the reference and the test probabilities match, PEM will be small.

During the transition period between phones, the phone identity is ambiguous, resulting in unreliable PPG values at the phone boundaries. This ambiguity is even more in singing compared to speech, because all phones are not always prominently articulated in singing, such as the word-end consonants [11], causing unclear boundaries. So to avoid the unreliable boundary values, we compute the phone-level scores by using only the center frames. Empirically, we found that 58% of the center frames results in the lowest detection error rate.

In singing, the vowels are often stretched in time, thus occupying a larger proportion of the word than that in speech [34]. We consider this characteristic of singing in computing the word-level scores, by either giving equal weights to all the phone-level scores of the word, or by giving weights to them according to the percentage of frames occupied by the phone in the word. We observe that frame-weighting the phone-scores shows higher correlation with human judgment compared to equally weighting them, which is intuitively justified as the long duration vowels have more time to make an impression on the listener. We also found that the PPG values for the short duration consonants tend to have more errors than the long duration

vowel segments. Thus frame-weighting also reduces the sensitivity of the score to PPG errors.

4. Experiment

We now conduct experiments to validate the proposed pronunciation evaluation strategy for singing. We test our hypotheses that lexicon modification leads to better word boundary alignment, investigate the performance of the speech pronunciation evaluation methods for singing with this modification, and validate with human scores collected via crowd-sourcing platform.

4.1. Dataset

From the DAMP dataset, we selected 24 singers (13 female, 11 male) each singing one of 6 unique English popular songs: *Let it go, Lovefool, I dreamed a dream, When I was your man, and Stay*. According to the metadata provided in DAMP, the singers belonged to different language speaking zones of the world: 4 from JA (Japanese), 1 from ZH (Chinese), 2 from ES (Spanish), 1 from FR (French), and the rest from EN (English). Pronunciation errors were caused by L1-influence, or unfamiliar lyrics, or both, in both native English and non-native English singer renditions, but more in non-native singers. For more details, please refer to our published dataset[1].

Since the songs were sung on Smule's karaoke app Sing!, all the renditions of each song were time-aligned. Forced-alignment using ASR is known to work well for short utterances. So we split the renditions into shorter utterances of 5-10 seconds by marking the line boundaries of one good pronunciation rendition of each song, and aligning the rest with DTW. This resulted in a total of 666 short sung utterances.

4.2. Human Annotations

We obtained three types of human annotations for a subset or the whole of this dataset to validate our automated word alignment and pronunciation scoring: word boundary time markings, pronunciation judgments at song-level, and word-level.

4.2.1. Word boundary markings

For the word-alignment validation experiment (Section 4.3.1), we manually marked the word boundaries of 100 utterances from the well-sung, i.e. correct pronunciation renditions of 5 singers who belonged to the EN zone (20 utterances per singer).

4.2.2. Word-level pronunciation judgments

For validating word-level automatic scoring, we asked two university students fluent in English to listen to 10 sung utterances from 10 singers, (5 from EN zone, and 5 from non-EN zone) i.e. 100 utterances, and marked the words in the lyrics that are mispronounced, i.e. substituted, deleted, or new words inserted. This is a binary judgment per word where the marking '1' for a word is to indicate incorrect pronunciation, and '0' is for correct pronunciation. In this way, we obtained word-level ground-truth pronunciation judgments for 990 words.

For the template-dependent method of scoring, we obtained the word-level evaluation for 10 utterances from 5 more EN zone singers with good pronunciation, who were considered as the reference templates for this experiment.

4.2.3. Song-level pronunciation judgments

To validate song-level pronunciation scores, we wanted to collect reliable human song-level pronunciation judgments in a scalable way, by leveraging on a crowd-sourcing platform, Amazon mechanical turk (MTurk). A method of proving reliability of the MTurk data is to observe the correlation between the MTurk data and that from a laboratory-controlled experiment [21].

MTurk data reliability test: In [35], our task was to build an algorithm for automatic singing quality evaluation. We asked 5 professionally trained musicians to give singing quality assessment for various singing parameters including pronunciation for 20 singers on a likert scale of 5. So we obtained lab-controlled average pronunciation scores from this experiment.

Here, we conducted the same experiment on MTurk, where Human Intelligence Tasks (HIT) consisted of a song audio file, along with its lyrics, followed by a questionnaire. The questionnaire now included five additional questions to inquire about the judge's music experience and English speaking fluency. The music experience related questions asked about how many years of vocal training/musical instrument training/stage performance experience have they got, a short description about their music experience, and asking them to transcribe randomly chosen four musical notes. The English speaking fluency questions asked if they were native English speakers, and to rate their own English speaking fluency. Each of the 20 singers' songs were rated by at least 7 human judges.

A human rating was rejected if two out of the three music-related questions showed that they did not have any music experience, and if they were non-native English speakers with English speaking fluency below 4. We also rejected a judgment that marked the exact same rating for all the questions. This shows that the rater was not serious about the task. After this questionnaire-based data clean-up procedure, we had at least 5 ratings per song, from which we computed the average ratings for each of the singing parameters. The average Pearson's correlation between these ratings and that from the controlled-lab experiment done by professional musician was 0.86. Thus the questionnaire-based data clean-up results in high correlation between the lab-controlled experiment and the MTurk experiment validating our hypothesis that we can get reliable subjective ratings for singing evaluation parameters, including pronunciation, from crowd-sourcing platforms.

We then implemented the same MTurk experiment for our new dataset of 24 songs (see Section 4.1). We have focused only on the pronunciation ratings for this paper. After the questionnaire-based data clean-up, the average inter-rating correlation for pronunciation between the 5 selected judges is 0.60. The average of the 5 ratings for each song is considered as the human ground-truth pronunciation score of the song.

4.3. Singing Pronunciation Evaluation Validation

In [26], we adapted baseline speech acoustic models (tri-phone HMM model trained on Librispeech corpus [36] using MFCC features) to singing with sung utterances from the DAMP dataset that resulted in WER of 36.32% from SAT+DNN models in an open loop decoding experiment. To account for the long duration vowels and obtain better singing-adapted models for alignment, we repeat the same experiment by using the singing-specific modified lexicon discussed in Section 3.1.1, which reduces the WER to 29.65% with SAT+DNN models. We also verified that our lexicon modification helps in modeling the long duration vowels (Table 1), i.e. longer duration vowels are modeled by more vowel repetitions in the lexicon. We use these singing-adapted models in the two-stage approach of pronunciation evaluation, as discussed in Section 3.

4.3.1. Word Alignment Validation

We validate the quality of word alignment from the forced-alignment of the singing-adapted SAT models with the sung utterance by comparing with the human annotations for word

[1]Dataset available here: `https://drive.google.com/open?id=19JPEWSBAM0ssatjBIJzAzjClxi2abt8w`

Table 1: *Effect of lexicon modification: # of vowels modeled by the different optional repetition variants in the lexicon, and the avg. duration of those vowels. (across the 666 sung utterances)*

repetition times in lexicon→	**0**	**1**	**2**	**3**
# of vowels	5804	3886	1299	402
avg. dur. of vowels (seconds)	0.218	0.380	0.674	1.518

Table 2: *Word alignment validation of well-sung renditions: the number of words within a range of absolute deviation of the automatic boundaries from the ground-truth. Total number of words=896. LEX: lexicon modification.*

Singing-adapted SAT Models	**<20ms**	**20-50ms**	**50-100ms**	**100-200ms**	**>200ms**
w/o LEX	635	115	82	24	40
LEX	748	74	33	9	32

boundaries. We also compare the word alignment performance of the lexicon-modified singing-adapted models with the one with the baseline singing-adapted models [26]. We expect our model to detect the word boundaries accurately in the sung utterances with good pronunciation, and that the word boundary detection should improve with the lexicon-modification. We compute the sum of absolute deviation of the start and the end boundaries from the ground-truth markings for every word as a measure of boundary deviation. Table 2 shows the number of words within different ranges of boundary deviations using SAT models with and without the lexicon modification, for sung utterances with good pronunciation. We see that lexicon modification improves the boundary alignment performance from 83.7% to 91.7% within 50ms of absolute boundary deviation, which is an 8% improvement.

4.3.2. Scoring Validation

We performed two experiments for validating our pronunciation scores for singing: word-level and song-level. In word-level evaluation, we compared the PPG-based template dependent and independent methods of scoring the aligned words, with the human judgments (see Section 3.2 and Section 4.2.2). In song-level evaluation, we compared the overall pronunciation score for a song rendition, with the human judgments obtained from MTurk (Section 4.2.3).

(1) Word-level scoring validation:

We wanted to see whether our scoring algorithms are able to correctly detect mispronounced words in a sung utterance. Table 3 compares the template dependent and independent methods of scoring with human word-level scores. It also shows the effect of lexicon modification on scoring. To evaluate the performance of the methods, we compute the metrics precision (Pre), recall (Rec), and F-score (F), where TP (True Positive) is the # of mispronounced words detected as mispronounced, TN (True Negative) is the # of correctly pronounced words detected as correctly pronounced, FP (False Positive) is the # of correctly pronounced words detected as mispronounced, and FN (False Negative) is the # of mispronounced words detected as correctly pronounced.

Template independent outperforms template-dependent method, with an equal error rate (EER) of 0.28 and accuracy of 0.72, compared to 0.47 and 0.52 respectively from template-dependent method. The main reason for high error rate is the high false positives in both the methods, but more in the template-dependent method. The template-independent method depends on the test utterance PPG and the lyrics, whereas the template-dependent method relies on the PPG of the test as well as the reference utterances. However the imperfect singing acoustic models that estimate the PPGs may cause errors in the PPGs. This results in false positives in both the methods, but more so in the template-dependent method because of errors in the reference template PPG.

Table 3: *Word-level scoring: Performance of automatic mispronunciation detection for singing and speech. P: Precision = TP/(TP+FP); R: Recall = TP/(TP+FN); F: F-score = 2.P.R/(P+R); FPR: False Positive Rate = FP/(FP+TN); FNR: False Negative Rate = FN/(FN+TP); Total number of words=990. LEX: lexicon modification.*

Method: Template-	**TP**	**TN**	**FP**	**FN**	**Pre**	**Rec**	**F**	**Acc**	**FPR**	**FNR**
Dependent (LEX)	71	445	410	64	0.15	0.53	0.23	0.52	0.48	0.47
Independent (LEX)	97	613	242	38	0.29	0.72	0.41	**0.72**	**0.28**	**0.28**
Independent (w/o LEX)	95	598	257	40	0.27	0.70	0.39	0.70	0.30	0.30

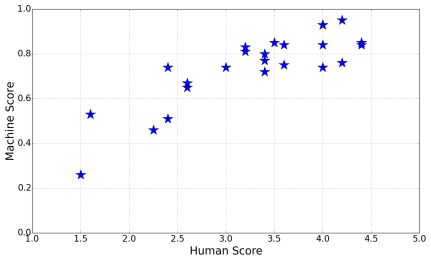

Figure 1: *Song-level score comparison: machine vs. humans. Pearson's correlation is 0.85.*

Also, with lexicon modification, template independent method performs better than without the modification. Mc Nemar's test P [37] is 0.038, implying the observed difference between the performance of the two algorithms would arise by chance on only 3.8% of occasions. So there is evidence of a statistically significant improvement in pronunciation evaluation performance based on the lexicon-modification method.

This experiment verified our hypothesis that the speech pronunciation evaluation methods are applicable for singing with singing-specific modifications.

(2) Song-level scoring validation:

We wanted to see if the word-level scores across all the utterances of a song can give an overall song-level pronunciation score that correlates with the human judgments. We computed the percentage of words detected as incorrectly pronounced ($\%error$) by template-independent method across all the utterances of a song by a singer, thus $1 - \%error$ is the measure for song-level pronunciation accuracy of a singer. The Pearson's correlation between the automatic and the average human annotated song-level pronunciation scores for the 24 songs is 0.85 (Figure 1). This verifies that the evaluation of pronunciation of singing based on template-independent pronunciation evaluation method gives reliable song-level assessment. We also found that computing the song-level scores with only the well-aligned words (aligned within 50 ms of the ground-truth boundaries) results in an even better correlation of 0.91 with the human scores. This means that better alignment leads to better evaluation performance.

5. Conclusions

We developed a strategy to compute reliable pronunciation evaluation scores for singing. We showed that duration-based lexicon modification for singing acoustic model adaptation results in improvement in word alignment as well as scoring accuracy. We also found that the template independent method of scoring with singing-specific modifications shows high correlation with human judgments both at word- and song-levels. Additionally, we verified that the subjective pronunciation scores for singing, that is needed for algorithm validation, can be reliably obtained through crowd-sourcing. Future work will involve analyzing the relationships between singer geographical origin, song difficulty level, and evaluation accuracy.

6. References

[1] S. O. Ali and Z. F. Peynircioğlu, "Songs and emotions: are lyrics and melodies equal partners?" *Psychology of Music*, vol. 34, no. 4, pp. 511–534, 2006.

[2] E. Brattico, V. Alluri, B. Bogert, T. Jacobsen, N. Vartiainen, S. Nieminen, and M. Tervaniemi, "A functional mri study of happy and sad emotions in music with and without lyrics," *Frontiers in psychology*, vol. 2, 2011.

[3] H. Nakata and L. Shockey, "The effect of singing on improving syllabic pronunciation – vowel epenthesis in japanese."

[4] A. J. Good, F. A. Russo, and J. Sullivan, "The efficacy of singing in foreign-language learning," *Psychology of Music*, vol. 43, no. 5, pp. 627–640, 2015.

[5] A. Norton, L. Zipse, S. Marchina, and G. Schlaug, "Melodic intonation therapy," *Annals of the New York Academy of Sciences*, vol. 1169, no. 1, pp. 431–436, 2009.

[6] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," *Speech Communication*, vol. 30, no. 2, pp. 83–93, 2000.

[7] Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr, and S. Yoon, "Accent detection and speech recognition for shanghai-accented mandarin." in *Interspeech*. Citeseer, 2005, pp. 217–220.

[8] H. Fujihara and M. Goto, "Lyrics-to-audio alignment and its application," in *Dagstuhl Follow-Ups*, vol. 3. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

[9] A. Loscos, P. Cano, and J. Bonada, "Low-delay singing voice alignment to text." in *ICMC*, 1999.

[10] N. F. Chen and H. Li, "Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*. IEEE, 2016, pp. 1–7.

[11] C. Gupta, D. Grunberg, P. Rao, and Y. Wang, "Towards automatic mispronunciation detection in singing," in *Proceedings of International Society for Music Information Retrieval (ISMIR), Suzhou, China, 2017*, 2017.

[12] P. Jha and P. Rao, "Assessing vowel quality for singing evaluation," in *National Conference on Communications (NCC) 2012, IEEE*, 2012, pp. 1–5.

[13] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[14] Y. Kim, H. Franco, and L. Neumeyer, "Automatic pronunciation scoring of specific phone segments for language instruction," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[15] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 398–403.

[16] A. Lee and J. Glass, "A comparison-based approach to mispronunciation detection," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 382–387.

[17] A. Lee, Y. Zhang, and J. Glass, "Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8227–8231.

[18] Smule, "Digital Archive Mobile Performances (DAMP)," https://ccrma.stanford.edu/damp/, [Online; accessed 15-March-2018].

[19] M. Marge, S. Banerjee, and A. I. Rudnicky, "Using the amazon mechanical turk for transcription of spoken language," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 5270–5273.

[20] F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer, "Crowdmos: An approach for crowdsourcing mean opinion score studies," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2416–2419.

[21] B. Naderi, T. Polzehl, I. Wechsung, F. Köster, and S. Möller, "Effect of trapping questions on the reliability of speech quality judgments in a crowdsourcing paradigm," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[22] J. Parson, D. Braga, M. Tjalve, and J. Oh, "Evaluating voice quality and speech synthesis using crowdsourcing," in *International Conference on Text, Speech and Dialogue*. Springer, 2013, pp. 233–240.

[23] H. Wang and H. Meng, "Deriving perceptual gradation of l2 english mispronunciations using crowdsourcing and the workerrank algorithm," in *Speech Database and Assessments (Oriental COCOSDA), 2012 International Conference on*. IEEE, 2012, pp. 145–150.

[24] A. Mesaros and T. Virtanen, "Automatic alignment of music audio and lyrics," in *Proceedings of the 11th Int. Conference on Digital Audio Effects (DAFx-08)*, 2008.

[25] ——, "Automatic recognition of lyrics in singing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 1, p. 546047, 2010.

[26] C. Gupta, H. Li, and Y. Wang, "Automatic alignment of lyrics to solo singing," in *submitted to ISMIR 2018*.

[27] R. J. Zatorre and S. R. Baum, "Musical melody and speech intonation: Singing a different tune," *PLoS biology*, vol. 10, no. 7, p. e1001372, 2012.

[28] S. Zhang, R. C. Repetto, and X. Serra, "Study of the similarity between linguistic tones and melodic pitch contours in beijing opera singing." in *ISMIR*, 2014, pp. 343–348.

[29] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices," in *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*. IEEE, 2007, pp. 215–218.

[30] T. L. New, M. Dong, P. Chan, X. Wang, B. Ma, and H. Li, "Voice conversion: From spoken vowels to singing vowels," in *Multimedia and Expo (ICME), 2010 IEEE International Conference on*. IEEE, 2010, pp. 1421–1426.

[31] J. Pons, R. Gong, and X. Serra, "Score-informed syllable segmentation for a cappella singing voice with convolutional neural networks," in *ISMIR 2017, Suzhou, China*, 2017.

[32] R. Gong, P. Cuvillier, N. Obin, and A. Cont, "Real-time audio-to-score alignment of singing voice based on melody and lyric information," in *Interspeech*, 2015.

[33] A. M. Harrison, W.-K. Lo, X.-j. Qian, and H. Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training," in *International Workshop on Speech and Language Technology in Education*, 2009.

[34] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, "The nus sung and spoken lyrics corpus: A quantitative comparison of singing and speech," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*. IEEE, 2013, pp. 1–9.

[35] C. Gupta, H. Li, and Y. Wang, "Perceptual evaluation of singing quality," in *Proceedings of APSIPA Annual Summit and Conference*, vol. 2017, 2017, pp. 12–15.

[36] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*, 2015.

[37] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*. IEEE, 1989, pp. 532–535.