TOWARDS AUTOMATIC MISPRONUNCIATION DETECTION IN SINGING

Chitralekha Gupta[1,2] **David Grunberg**[1] **Preeti Rao**[3] **Ye Wang**[1]

[1] School of Computing, National University of Singapore, Singapore
[2] NUS Graduate School for Integrative Sciences and Engineering,
National University of Singapore, Singapore
[3] Department of Electrical Engineering, Indian Institute of Technology Bombay, India
chitralekha@u.nus.edu, wangye@comp.nus.edu.sg

ABSTRACT

A tool for automatic pronunciation evaluation of singing is desirable for those learning a second language. However, efforts to obtain pronunciation rules for such a tool have been hindered by a lack of data; while many spoken-word datasets exist that can be used in developing the tool, there are relatively few sung-lyrics datasets for such a purpose. In this paper, we demonstrate a proof-of-principle for automatic pronunciation evaluation in singing using a knowledge-based approach with limited data in an automatic speech recognition (ASR) framework. To demonstrate our approach, we derive mispronunciation rules specific to South-East Asian English accents in singing based on a comparative study of the pronunciation error patterns in singing versus speech. Using training data restricted to American English speech, we evaluate different methods involving the deduced L1-specific (native language) rules for singing. In the absence of L1 phone models, we incorporate the derived pronunciation variations in the ASR framework via a novel approach that combines acoustic models for sub-phonetic segments to represent the missing L1 phones. The word-level assessment achieved by the system on singing and speech is similar, indicating that it is a promising scheme for realizing a full-fledged pronunciation evaluation system for singing in future.

1. INTRODUCTION

Educators recommend singing as a fun and effective language learning aid [6]. In fact, it has been observed that the use of songs and karaoke is helpful in teaching and improving pronunciation in adult second language (L2) classes [1, 17] . Scientific studies have shown that there is a connection between the ability of phonemic production of a foreign language and singing ability [16], and singing ability often leads to better imitation of phrases in an unknown language [15]. More recently, evidence from experimental psychology suggests that learning a new language through songs helps improve vocabulary gain, memory recall, and pronunciation [11]. Additionally, singing releases the need to focus on prosody, as melody of the song overrides the prosodic contrasts while singing [13]. So, given a familiar melody, all the attention can be on articulating the lyrics correctly.

Given the potential of singing in pronunciation training, it is of interest to research automatic pronunciation evaluation for sung lyrics similar to the large body of work in computer-aided pronunciation training (CAPT) for speech [18, 29]. There is little work on mispronunciation detection for sung lyrics. Jha et al. attempted to build a system for evaluating vowels in singing with Gaussian Mixture Model (GMM) and linear regression using Mel Frequency Cepstral Coefficients (MFCC) and pitch as features [12]. However, they did not account for possible pronunciation error patterns in singing, and further, their work did not extend to consonants. There have been a few other studies that have subjectively compared the pronunciation in singing versus that in speech. Yoshida et al. [26] conducted a subjective mispronunciation analysis in singing and speech in English for Japanese natives and found that the subjects familiar with singing tend to make less mistakes in pronunciation while singing than speaking. Another study found that the most frequent pronunciation errors by Indonesian singers in singing English songs occur in the consonants [21]. None of these studies however attempted to build an automatic evaluator of pronunciation in singing.

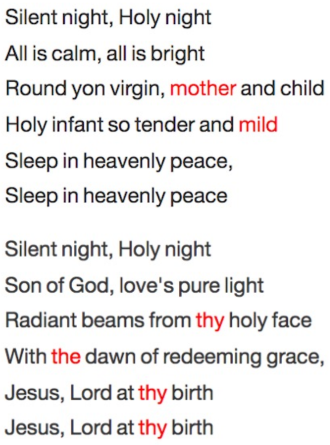
Though studies have been conducted to compare singing and speech utterances [5, 19], the automated assessment of singing pronunciation is hampered by the lack of training datasets of phone-level annotated singing. Duan et al. created a dataset to analyse the similarities and differences between spoken and sung phonemes [8]. This dataset consists of sung and spoken utterances from 12 unique subjects, out of which 8 were noted as non-native speakers. But their work did not study the pronunciation error patterns in singing or speech. A part of this dataset was phonetically transcribed in 39 CMU phones [25], which is inadequate for annotating non-native pronunciations. We use a subset of audio clips from this dataset for our work (as explained in Section 2.2). But we did not use their phonetic annotations due to these limitations.

In this work, we demonstrate a knowledge-based approach with limited data to automatically evaluate pronunciation in singing in an automatic speech recognition (ASR) framework. We will adopt a basic method of phonetic evaluation that is used for speech, i.e. deriving pronunciation variants based on L1-L2 pair error patterns, and incorporating this knowledge in the ASR framework for evaluation [3, 23]. In our study, we analyze the error patterns in singing versus those in speech in the accents of South-East Asia - Malaysian, Indonesian, and Singaporean. South-East Asia is one of the most populous regions of the world, where the importance of speaking standard English has been recognized [14], and hence such a pronunciation evaluation system is desired. Given that the data available to train acoustic models is restricted to a native American English speech database [10], we present a novel approach of combining sub-phonetic segments to represent missing L1-phones. Also, we demonstrate how the systematic incorporation of the knowledge of the error patterns helps us obtain a reliable pronunciation evaluation system for non-native singing.

2. PRONUNCIATION ERROR PATTERNS

2.1 Previous Studies

In the process of learning a second language L2, a common type of mispronunciation is replacing phones of L2 that do not exist in the native language (L1) with the closest sounding phoneme of L1 [4]. In Malaysian, Singaporean, and Indonesian English, the dental fricatives /th/ and /dh/ are often substituted by alveolar stops /t/ and /d/ respectively (eg. “three”→“tree”, “then”→“den”). These accents are influenced by Malay, Mandarin, and Indonesian languages, in which the dental fricatives /th/ and /dh/ are absent [2, 7, 9]. Also, a pattern particularly seen in Indonesian English accent is that the alveolar stop consonants /t/ and /d/ tend to be substituted by their apico-dental unaspirated stop variant. The reason for this confusion is that in the Indonesian language, the phones /d/ and /t/ can be both alveolar or dental [2, 22]. Another pattern in Singaporean and Malaysian accents is that they tend to omit word-end consonants, or replace them with glottal stops. Note the lack of word-end consonants in the Malay counterparts of words like “product” is “produk”. Also in Mandarin, most words do not end with a consonant, except /ng/ and /n/. Vowel difficulties are seen in all these accents, such as long-short vowel confusions like “bead”→“bid”, because the long /iy/ is absent in the Indonesian language. Another clear pattern reported is the voiced post-alveolar approximant /r/ in English being pronounced as an apical post-dental trill /r/ in Indonesian, that sounds like a rolling “r”.



Silent night, Holy night
All is calm, all is bright
Round yon virgin, mother and child
Holy infant so tender and mild
Sleep in heavenly peace,
Sleep in heavenly peace

Silent night, Holy night
Son of God, love's pure light
Radiant beams from thy holy face
With the dawn of redeeming grace,
Jesus, Lord at thy birth
Jesus, Lord at thy birth

Figure 1: Example of word-level subjective evaluation on the website (incorrect words marked in red).

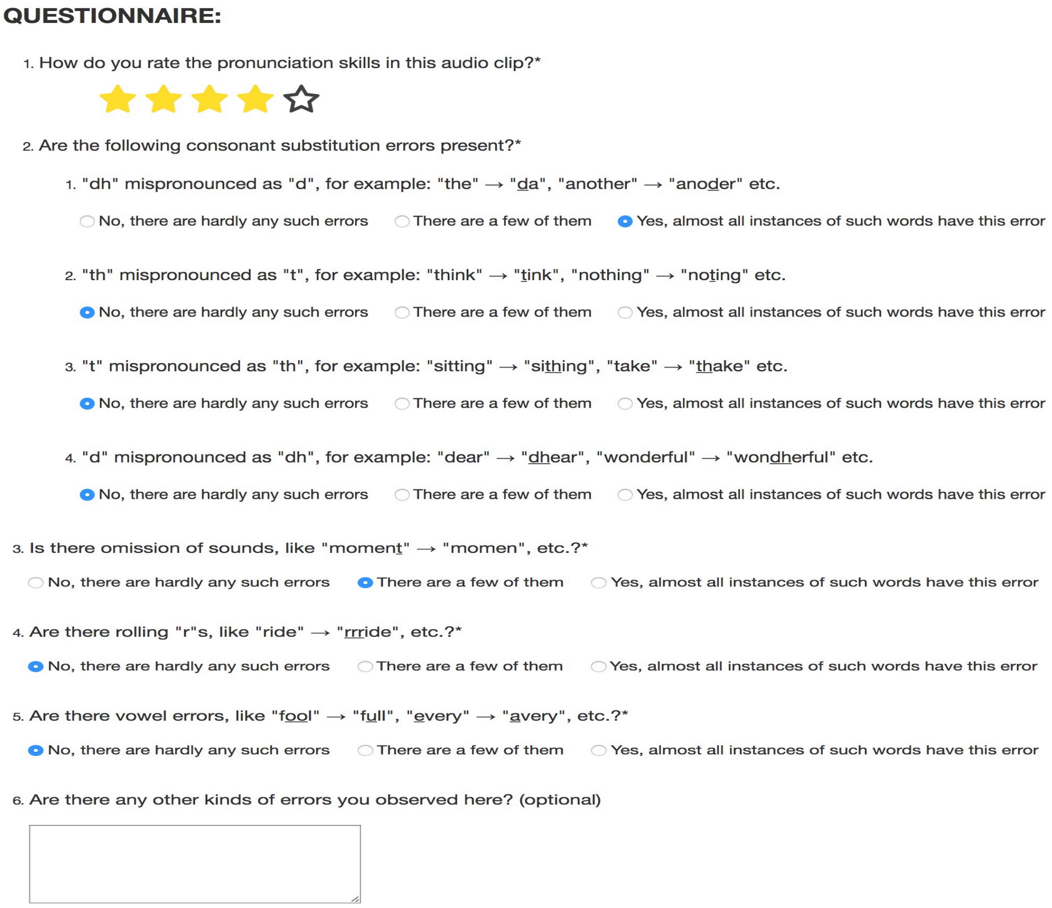
Here, we investigate the general rules of mispronunciation in singing, which will be then used for automatic pronunciation evaluation in singing. We will, henceforth, refer to the L1 roots of Malaysian, Singaporean, and Indonesian English as M, S, and I, respectively.

2.2 Dataset

The dataset (a subset of a published dataset [8]) consists of a total of 52 audio files: 26 sung and 26 spoken, from 15 popular English songs (like Do Re Mi, Silent Night, etc.). Each song has 28 lines (phrases) on an average. These songs were sung and spoken by 8 unique subjects (4 male, 4 female) - 3 Indonesian, 3 Singaporean, and 2 Malaysian. The subjects were students at National University of Singapore, with experience in singing. The subjects were asked to familiarize themselves with lyrics of the songs before the recording session and could use a printed version of the lyrics for their reference during recording. No background accompaniments were used while recording except for metronome beats which were sent to headphones.

We developed a website to collect subjective ratings for this dataset. The website consisted of the audio tracks, their corresponding lyrics, and a questionnaire. Each word in the lyrics could be clicked by the rater to mark it as incorrectly pronounced (red), as shown in the screenshot of the webpage in Figure 1. For each sung and spoken audio clip, the raters were asked to first listen to the track and mark the words in the lyrics that were incorrectly pronounced, and then fill up the questionnaire based on their word-error judgment, as shown in Figure 2. We asked 3 human judges (two North American native English speakers, and one non-native speaker proficient in English), to do this task. Here, native English pronunciation (North American) is considered as the benchmark for evaluating pronunciation.

In the questionnaire, the judges evaluated the overall pronunciation quality on a 5 point scale. On a 3 point scale, they evaluated the presence of each consonant substitution (C1-C4), vowel replacement (V), word-end consonant deletion (CD), and rolling “r” (R), each corresponding to the rules listed in Table 1, where rating 1 means there



QUESTIONNAIRE:

1. How do you rate the pronunciation skills in this audio clip?*

★★★★☆

2. Are the following consonant substitution errors present?*

1. "dh" mispronounced as "d", for example: "the" → "da", "another" → "anoder" etc.

○ No, there are hardly any such errors ○ There are a few of them ● Yes, almost all instances of such words have this error

2. "th" mispronounced as "t", for example: "think" → "tink", "nothing" → "noting" etc.

● No, there are hardly any such errors ○ There are a few of them ○ Yes, almost all instances of such words have this error

3. "t" mispronounced as "th", for example: "sitting" → "sithing", "take" → "thake" etc.

● No, there are hardly any such errors ○ There are a few of them ○ Yes, almost all instances of such words have this error

4. "d" mispronounced as "dh", for example: "dear" → "dhear", "wonderful" → "wondherful" etc.

● No, there are hardly any such errors ○ There are a few of them ○ Yes, almost all instances of such words have this error

3. Is there omission of sounds, like "moment" → "momen", etc.?*

○ No, there are hardly any such errors ● There are a few of them ○ Yes, almost all instances of such words have this error

4. Are there rolling "r"s, like "ride" → "rrride", etc.?*

● No, there are hardly any such errors ○ There are a few of them ○ Yes, almost all instances of such words have this error

5. Are there vowel errors, like "fool" → "full", "every" → "avery", etc.?*

● No, there are hardly any such errors ○ There are a few of them ○ Yes, almost all instances of such words have this error

6. Are there any other kinds of errors you observed here? (optional)

Figure 2: Questionnaire for every audio clip on the subjective evaluation webpage.

are hardly any errors of that category, while rating 3 means almost all of the occurrences have error. We shall refer to these ratings as Rating Set 1. These questions cover all the phone error categories in speech in the accents concerned according to the literature, as described in section 2.1. Additionally, the questionnaire included a comment text-box in which the rater could mention about any other kinds of errors that they observed, which were not covered by the other questions. In this way, we tried to ensure that the subjective evaluation was not biased by the mentioned error categories in the questionnaire.

The average inter-judge correlation (Pearson's) of the overall rating question was 0.68 for the sung clips and 0.62 for the spoken clips, and that for the questions on error-categories was 0.89 for the sung clips and 0.74 for the spoken clips. Thus the inter-judge agreement was high. Also, in the comment text-box, the judges provided only rare minor comments, such as mispronouncing "want to" as "wanna", which could not be categorized as systematic errors due to L1 influence, and hence are not included in the current study.

We chose the word-level pronunciation assessment ("correct"/ "wrong") of one of the North American native English speaking judges as the ground truth. We shall refer to these ratings as Rating Set 2.

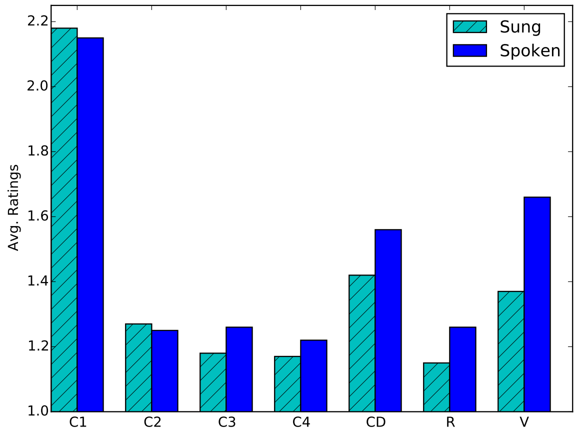


Figure 3: Average subjective rating for seriousness of each error category for singing and speech. Error category labels are as per Table 1.

2.3 Analysis of Error Patterns: Singing vs. Speech

From the rating set 2, we obtained a list of consonant and vowel error patterns in speech and singing, and examples of such words, as shown in Table 2. These error categories can be directly mapped to the questions in the questionnaire, and are consistent with the literature on error patterns in South-East Asian accents.

Our aim here is to derive a list of rules relevant to mis-

Label	Error Rule	p-value
C1	/dh/ → /d/ (WB,WM)	0.414
C2	/th/ → /t/ (WB,WM)	0.382
C3	/t/ → /th/ (WB,WM,WE)	0.079
C4	/d/ → /dh/ (WB,WM,WE)	0.243
CD	Consonant deletion (WE)	**0.032**
R	Rolling "r"	0.112
V	Vowel substitutions	**$5*10^{-5}$**

Table 1: Statistical significance of the difference of each error category between singing and speech (WB: Word Beginning, WM: Word Middle, WE: Word Ending).

Consonants			
Error	**WB**	**WM**	**WE**
/dh/ → /d/	the, they, thy, then	mother, another,	
/th/ → /t/	thought, thread,	nothing	
/t/ → /th/	to, tea, take	spirits, into, sitting	note, it, got
/d/ → /dh/	drink, dear	outdoors, wonderful	cloud, world
Consonant deletion (only WE)			night, moment, child
rolling "r"	run, ray,round	bread, drop, bright	brighter, after

Vowels		
Error	**Actual word**	**What is spoken**
ow-->ao	golden	gawlden
uw-->uh	fool	full
iy-->ih,ix	seem, sees, sleeping,	sim, sis, slipping
eh-->ae	every	avry

Table 2: Error categories in singing and speech, and examples of words where they occur.

L1	Label	Rule	Dictionary A		Dictionary B	
			Can.	Mis.	Can.	Mis.
M, S, I	C1	WB,WM /dh/ → /d/	dh → vcl d		dh → vcl d vcl dh →vcl d	
M, S, I	C2	WB,WM / th/ → /t/	th → cl t		th → cl t cl th → cl t	
I	C3	WB,WM,WE /t/ → /th/	cl t → th		cl t → cl th	
I	C4	WB,WM,WE /d/ → /dh/	vcl d → dh		vcl d→ vcl dh	

Table 3: Mispronunciation rules for singing, and corresponding transcriptions for Dictionaries A and B. Can.: Canonical, Mis.: Mispronunciation (cl: unvoiced closure, vcl: voiced closure, dh: dental voiced fricative, d: alveolar voiced stop, th: dental unvoiced fricative, t: alveolar unvoiced stop).

pronunciation in singing. From rating set 1, we compute the average subjective rating for each of the mispronunciation rules for singing and speech, as shown in Figure 3. To identify the rules that are relevant for singing, we compute the difference of the average ratings between singing and speech for every rule for each of the 26 pairs of audio files, and compute the $p - value$ of this difference. For a particular rule, if the overall average rating of singing is less than that of speech, and the difference of average ratings between singing and speech is significant ($p - value < 0.05$), then that particular kind of mispronunciation is not frequent in singing, and thus the rule is not relevant for singing. For example, we found that most of the detectable mispronunciations in singing were seen in consonants, which agrees with the literature previously discussed [2, 7, 9, 21, 22]. The mean rating for the question "Are there vowel errors?" was much lower for singing than for speech, meaning that there are fewer vowel errors perceived in singing than in speech (as shown in Figure 3). The difference of the 26 ratings for this question between singing and speech is statistically significant ($p - value = 5 \times 10^{-5}$) (Table 1), and hence confirms this trend. In singing, the vowel duration and pitch in singing are usually dictated by the corresponding melodic note attributes, which makes it different from spoken vowels. For example, the word "sleep" is stretched in duration in the song "Silent Night", thus improving the pronunciation of the vowel. However, in speech the word might tend to be pronounced as "slip". The explanation could lie in the way singers imitate vowels based on timbre (both quality and duration) rather than by the categorical mode of speech perception which is applied only to the consonants. In the same way, we also found that the "word-end consonant deletion" category of errors is significantly less frequent in singing than in speech ($p - value = 0.032$) (Table 1). This implies that either the word-end stop consonants like /t/ and /d/ are hardly ever omitted in singing or are imperceptible to humans. This leads us to the conclusion that only a subset of the error patterns that occur in speech are seen to be occurring in singing. This is a key insight that suggests a possible learning strategy: learning this "subset" of phoneme pronunciation through singing, and the rest through speech.

Another interesting inference from Figure 3 is that on an average, singing has a lower extent of perceived pronunciation errors compared to speech, which is also indicated by the average of the overall rating, which is higher for singing (singing = 3.87, speech = 3.80). This suggests that if the non-native subject is familiar with a song and its lyrics, he/she makes fewer pronunciation mistakes in singing compared to speech. Also, a non-native speaking accent is typically characterised by L1-influenced prosody as well such as stress and intonation aspects, which can influence subjective ratings. Singing on the other hand uses only the musical score and is therefore devoid of L1 prosody cues.

Table 3 lists the L1-specific mispronunciation rules for singing that we derived, in which the word-end consonant deletion and vowel substitution rules have been omitted for reasons mentioned above. In the Indonesian accent, the phone "r" was often replaced with a rolling "r" (trill) (Figure 3), which occurs frequently in singing as well (Table 1). But this phone is absent in American English, so we do not have a phonetic model to evaluate it. So, we have excluded this error pattern in this study.

With our dataset of sung utterances from 8 subjects, we could see clear and consistent mispronunciation patterns across the speakers in our subjective assessment study, and these patterns agree with the phonetic studies of L1 influence on English from these accents in the literature. Therefore, even if the dataset is small, it captures all the expected diversity.

3. AUTOMATIC DETECTION OF MISPRONUNCIATION IN SINGING

Our goal is to use an automatic speech recognition (ASR) system to detect mispronunciations in singing. In previous studies, vowel error patterns in Dutch pronunciation of L2 learners were used by Doremalen et al. to improve automatic pronunciation error detection systems [23]. In another work, Black et al. derived common error patterns in children's speech and used an adapted lexicon in an extended decoding tree to obtain word-level pronunciation assessment [3]. A standard way to detect mispronunciation is to let the ASR recognize the most likely sequence of phonemes for a word from a given set of acceptable (canonical) and unacceptable (mispronounced) pronunciation variants of that word. A pronunciation is detected as unacceptable if the chosen sequence of phonemes belongs to the list of unacceptable pronunciation variants of the word (called the "lexicon" or the "LEX" method in [3]). While the present work is similar in principle to the above, we face the additional challenge of lack of training data for the L1 phones not present in L2. Yu et al. [27] have used a data-driven approach to convert the foreign-language lexicon (L2) to native-language lexicon (i.e. using L1 phones only), where they had large L1 speech training data, in contrast to our case of availability of L2 training speech only. In the current work, we use a novel knowledge-based approach to overcome the constraints of lack of L1 training data for both speech and singing. We compare the case of restricting ourselves to L2 phones with a method that uses L1 phones derived from a combination of sub-phonetic segments of L2 speech to approximate unavailable L1 phones.

We compare a Dictionary A that contains only American English (TIMIT [10]) phones (L2), with a Dictionary B that contains TIMIT phones+modified (L1-adapted) phones. To design Dictionary B, we compared the phones of the South-East Asian accents with that of the American English TIMIT speech dataset [10]. We found that the dental fricatives /th/ and /dh/ are often mispronounced as alveolar stops /t/ and /d/ respectively (rules C1, C2). Both of the substituted phones /t/ and /d/ are present in American English, and hence their phone models are available in TIMIT. But when L1 is Indonesian, the alveolar stop consonants /t/ and /d/ tend to be substituted by their apico-dental unaspirated stop variant (rules C3, C4), as explained in Section 2.1. But dental stop phones are not annotated in American English datasets like TIMIT [10]. In order to solve this problem of lack of dental stop phone models in L2, we combined sub-phonetic TIMIT models. We observed that the dental stop phones consist of a closure period followed by a burst period with dental place of articulation. So we combined the TIMIT models for unvoiced closure model /cl/ with the unvoiced dental fricative model /th/ to approximate unvoiced dental stop /t/, as shown in Figure 4, and voiced closure model /vcl/ with the voiced dental fricative model /dh/ to obtain voiced dental stop /d/. It is important to note that in these accents, the dental fricatives /th/ and /dh/ are also often substituted by dental stops /cl th/ and /vcl dh/. But this particular substitution pattern is common in American English [28], and hence not considered to be mispronunciation. Hence, we add these variants to the list of acceptable variants (canonical).

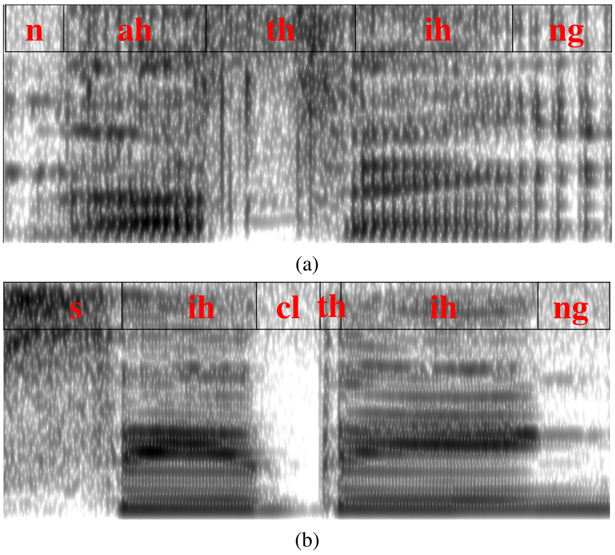


Figure 4: (a) American speaker (TIMIT) articulating word-middle unvoiced dental fricative /th/ in "nothing" (note: there is no closure) (b) Indonesian speaker substituting unvoiced alveolar stop with unvoiced dental stop ("sitting" as "sithing") modeled as /cl th/.

In summary, the mispronunciation rules in Dictionary B are: dental fricative and stop /dh/ being mispronounced as alveolar stop /d/ (L1: M, S, I); dental fricative and stop /th/ being mispronounced as alveolar stop /t/ (L1: M, S, I); alveolar stop /t/ being mispronounced as dental stop /th/ (L1: I); and alveolar stop /d/ being mispronounced as dental stop /dh/ (L1: I). These mispronunciation rules are listed in Table 3.

3.1 Methodology

We use the toolkit KALDI [20] for training 48 context independent GMM-HMM and DNN-HMM phonetic models using the TIMIT train set [10] with the parameters set by Vesely et al. [24]. The HMM topology is 3 active states, the MFCC features are frame-spliced by 11 frames, dimension-reduced by Linear Discriminant Analysis (LDA) to 40 dimensions. Maximum Likelihood Linear Transformation (MLLT), feature-space Maximum Likelihood Linear Regression (fMLLR), and Cepstral Mean and Variance Normalization (CMVN) are applied for speaker adaptive training. The DNN has 6 hidden layers, 2048 hidden units per layer. The Restricted Boltzmann Machine (RBM) pre-training algorithm is contrastive divergence and the frame cross-entropy training is done by mini-batch stochastic gradient descent. Phone recognition performance of the acoustic models trained and tested on TIMIT was consistent with the literature [24].

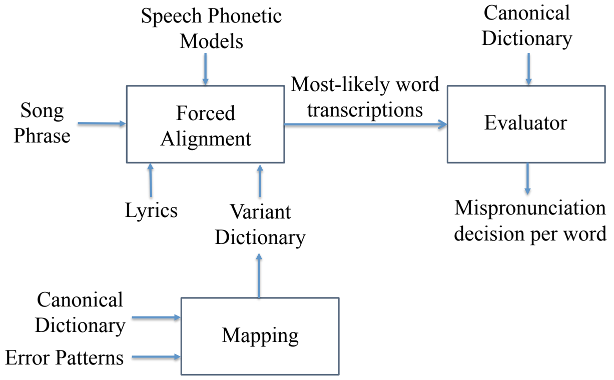


Figure 5: Overview of automatic mispronunciation detection in singing.

L1 (#EP-W)	AM	Speech						Singing					
		Dictionary A			Dictionary B			Dictionary A			Dictionary B		
		P	R	F	P	R	F	P	R	F	P	R	F
M,S **(245)**	**DNN-HMM**	0.51	0.58	0.54	0.60	0.68	**0.63**	0.62	0.65	0.63	0.69	0.66	**0.67**
	GMM-HMM	0.41	0.54	0.47	0.49	0.63	0.55	0.54	0.52	0.53	0.55	0.51	0.53
#GT-E		78						86					
I **(834)**	**DNN-HMM**	0.29	0.55	0.38	0.56	0.46	**0.50**	0.23	0.59	0.33	0.42	0.54	**0.47**
	GMM-HMM	0.27	0.50	0.35	0.46	0.40	0.43	0.21	0.58	0.31	0.33	0.34	0.34
#GT-E		219						176					

Table 4: 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); AM: Acoustic Models; #GT-E: no. of error-prone words mispronounced; #EP-W: no. of error-prone words. L1 languages - M: Malaysian, S: Singaporean, and I: Indonesian.

We use these speech trained acoustic-phonetic models, along with the L1-specific variant dictionary (A or B) to force-align the lyrics to the sung and spoken audio files, to obtain word-level automatic pronunciation evaluation by the "LEX" method, as described before. An overview of this system is shown in Figure 5. We first segment the audio files at phrase level by aligning its pitch track with a template containing a reference pitch track and marked phrase-level boundaries, using dynamic time warping. The 26 songs are segmented into 740 phrases, containing a total of 5107 words. For singing, out of these 5107 words, 1079 words are the error-prone words, i.e. they fall under the mispronunciation rules. Only 14 out of the rest 4028 non-error-prone words (0.3%) are subjectively evaluated as mispronounced in singing, which confirms that the mispronunciation rules for singing are correctly identified. To compare speech and singing, we apply the same rules for the speech phrases because we expect that the words that are mispronounced in singing are likely to be mispronounced in speech as well.

Table 4 shows the validation results for the L1-specific error-prone words from the two acoustic model configurations in singing and speech, using the dictionaries A and B, where the ground truth is the word-level subjective evaluation as obtained in rating set 2. To evaluate the performance of the system, we compute the metrics precision, recall, and F-score [3], where TP (True Positive) is the number of mispronounced words detected as mispronounced, FP (False Positive) is the number of correctly pronounced words detected as mispronounced, and FN (False Negative) is the number of mispronounced words detected as correctly pronounced (Table 4).

4. RESULTS AND DISCUSSION

We note that the method of combining sub-phonetic American English models for approximating the missing phone models of L1 is effective as the F-scores indicate that the system using dictionary B outperforms the one using A in all the cases. DNN-HMM outperforms GMM-HMM consistently for the task for pronunciation evaluation in singing, as it has been widely observed in speech recognition. Also, the F-score values of singing and speech are similar, which shows that our knowledge-based approach for singing pronunciation evaluation is promising.

A source of false positives is the rule /t/→/th/ which causes error when /t/ is preceded by a fricative (eg. /s/), for example "just" [jh, ah, s, cl, t]. Since both /s/ and /th/ are fricatives, the system gets confused and aligns /th/ at the location of /s/. A way to handle such errors is to obtain features specific to classifying the target and the competing phonemes, which will be explored in the future.

5. CONCLUSION

In this paper, we have analysed pronunciation error patterns in singing vs. those in speech, derived rules for pronunciation error patterns specific to singing, and demonstrated a knowledge-based approach with limited data towards automatic word-level assessment of pronunciation in singing in an ASR framework. From subjective evaluation of word pronunciation, we learn that nearly all identified mispronunciations have an L1-based justification, and singing has only a subset of the errors found in speech. We provide the rules that predict singing mispronunciations for a given L1. In order to solve the problem of unavailable L1 phones due to the lack of training speech data from L1 speakers, we propose a method that uses a combination of sub-phonetic segments drawn from the available native L2 speech to approximate the unavailable phone models. This method is shown to perform better than the one that restricts to only L2 phones. And finally, the performance of this system on singing and speech is comparable, indicating that this approach is a promising method for developing a full-fledged pronunciation evaluation system. In future, we would explore a combination of data-driven methods such as in [27] and our knowledge-based methods to improve the mispronunciation detection accuracy.

6. REFERENCES

[1] Developing pronunciation through songs. `https://www.teachingenglish.org.uk/article/developing-pronunciation-through-songs`. Accessed: 2017-04-27.

[2] B. Andi-Pallawa and A.F. Abdi Alam. A comparative analysis between English and Indonesian phonological systems. *International Journal of English Language Education*, 1(3):103–129, 2013.

[3] M. Black, J. Tepperman, and S. Narayanan. Automatic prediction of children's reading ability for high-level literacy assessment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):1015–1028, 2011.

[4] P. Bonaventura, D. Herron, and W. Menzel. Phonetic rules for diagnosis of pronunciation errors. In *KONVENS: Sprachkommunikation*, pages 225–230, 2000.

[5] W. Chou and G. Liang. Robust singing detection in speech/music discriminator design. In *IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings 2001*, pages 865–868, 2001.

[6] F. Dege and G. Schwarzer. The effect of a music program on phonological awareness in preschoolers. *Frontiers in Psychology*, 2(124):7–13, 2011.

[7] D. Deterding. *Singapore English*. Edinburgh University Press, 2007.

[8] Z. Duan, H. Fang, L. Bo, K. C. Sim, and Y. Wang. The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA) 2013, IEEE*, pages 1–9, 2013.

[9] S.Y. Enxhi, T.B. Hoon, and Y.M. Fung. Speech disfluencies and mispronunciations in English oral communication among Malaysian undergraduates. *International Journal of Applied Linguistics and English Literature*, 1(7):19–32, 2012.

[10] J. Garofolo. TIMIT acoustic-phonetic continuous speech corpus. In *Philadelphia: Linguistic Data Consortium*, 1993.

[11] A. Good, F. Russo, and J. Sullivan. The efficacy of singing in foreign-language learning. *Psychology of Music*, 43(5):627–640, 2015.

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

[13] I. Lehiste. Prosody in speech and singing. In *Speech Prosody 2004, International Conference*, 2004.

[14] L. Lim, A. Pakir, and L. Wee. *English in Singapore: Modernity and management*, volume 1. Hong Kong University Press, 2010.

[15] C. Markus and S.M. Reiterer. Song and speech: examining the link between singing talent and speech imitation ability. *Frontiers in psychology*, 4:874, 2013.

[16] R. Milovanov, P. Pietil, M. Tervaniemi, and P. Esquef. Foreign language pronunciation skills and musical aptitude:a study of Finnish adults with higher education. *Learning and Individual Differences*, 20(1):56–60, 2010.

[17] H. Nakata and L. Shockey. The effect of singing on improving syllabic pronunciation–vowel epenthesis in japanese. In *International Conference of Phonetic Sciences*, 2011.

[18] L. Neumeyer, H. Franco, V Digalakis, and M. Weintraub. Automatic scoring of pronunciation quality. *Speech Communication*, 30(2):83–93, 2000.

[19] Y. Ohishi, M. Goto, K. Itou, and K. Takeda. Discrimination between singing and speaking voices. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech 2005)*, pages 1141–1144, 2005.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[21] I. Riyani and J. Prayogo. An analysis of pronunciation errors made by Indonesian singers in Malang in singing English songs. *SKRIPSI Jurusan Sastra Inggris-Fakultas Sastra UM*, 2013.

[22] E. Setyati, S. Sumpeno, M. Purnomo, K. Mikami, M. Kakimoto, and K. Kondo. Phoneme-Viseme mapping for Indonesian language based on blend shape animation. *IAENG International Journal of Computer Science*, 42(3), 2015.

[23] J. van Doremalen, C. Cucchiarini, and H. Strik. Automatic pronunciation error detection in non-native speech: The case of vowel errors in Dutch. *The Journal of the Acoustical Society of America*, 134(2):1336–1347, 2013.

[24] K. Vesely, A. Ghoshal, L. Burget, and D. Povey. Sequence-discriminative training of deep neural networks. In *Interspeech*, pages 2345–2349, 2013.

[25] R. Weide. The Carnegie Mellon pronouncing dictionary [cmudict. 0.6], 2005.

[26] K. Yoshida, N. Takashi, and I. Akinori. Analysis of English pronunciation of singing voices sung by Japanese speakers. In *International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IEEE)*, 2014.

[27] D. Yu, L. Deng, P. Liu, J. Wu, Y. Gong, and A. Acero. Cross-lingual speech recognition under runtime resource constraints. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4193–4196. IEEE, 2009.

[28] S. Zhao. Stop-like modification of the dental fricative /dh/: An acoustic analysis. *The Journal of the Acoustical Society of America*, 128(4):2009–2020, 2010.

[29] 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*, pages 217–220. Citeseer, 2005.