Computational Pronunciation Analysis in Sung Utterances

1st Emir Demirel
Centre for Digital Music
Queen Mary University of London
London, UK
e.demirel@qmul.ac.uk

2nd Sven Ahlbäck
Doremir Music Research AB
Stockholm, Sweden
sven.ahlback@doremir.com

3rd Simon Dixon
Centre for Digital Music
Queen Mary University of London
London, UK
s.e.dixon@qmul.ac.uk

***Abstract*—Recent automatic lyrics transcription (ALT) approaches focus on building stronger acoustic models or in-domain language models, while the pronunciation aspect is seldom touched upon. This paper applies a novel computational analysis on the pronunciation variances in sung utterances and further proposes a new pronunciation model adapted for singing. The singing-adapted model is tested on multiple public datasets via word recognition experiments. It performs better than the standard speech dictionary in all settings reporting the best results on ALT in a capella recordings using n-gram language models. For reproducibility, we share the sentence-level annotations used in testing, providing a new benchmark evaluation set for ALT.**

***Index Terms*—automatic lyrics transcription, music information retrieval, computational linguistics, automatic speech recognition**

I. Introduction

The articulation of words during singing is influenced by the melodic line causing temporal variations in duration and the acoustic properties of the signal like pitch, timbre and loudness. Singers may even add an extra formant on top of the ones that characterize vowels, namely the singer's formant [1], increasing the perceived loudness of the voice. Consequently, these articulations during singing may alter the ways that words are pronounced and how they are perceived, thus affecting overall intelligibility [2]. Similarly, the performance of ALT systems that attempt to automatically recognize words from singing voice also gets affected by these altered pronunciations and variations in the acoustic properties. While only a little focus has been drawn to computationally model the pronunciation variances in singing performances, Gupta et al. [8] proposed to use a vowel-extended version of a standard lexicon with regards to the longer vowels in sung utterances and observed considerable improvement in word recognition.

In this study, we aim to shed a light on the pronunciation differences in sung utterances compared to speech by conducting a novel quantitative analysis on the phoneme level, identifying a number of systematic cases. Furthermore, we propose a new lexicon adaptation method for modelling of singing, and evaluate its effectiveness through word recognition rates over a number of open-source data sets. Additionally, we test the findings of the phonetic analysis through an error analysis. For reproducibility, we share the annotations publicly.

This paper begins by establishing a contextual ground to understand how the pronunciation models are employed in common hybrid-ASR frameworks. In Section 3, the details of the phonetic analysis are presented. Then, a method is proposed for extending a standard pronunciation dictionary for singing with respect to the observations of the analyses in Section 3. Section 4 explains the ALT setup for the recognition experiments. Section 5 provides the error analysis in terms of word and character error rates.

ED received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765068.

II. Related Work

The task of ASR can be summarized as finding the most probable word sequence, $\widehat{\mathbf{w}}$, given a sequence of acoustic observations, $\mathbf{X}$, which can be expressed using the following formula:

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}\, P(\mathbf{w}) \sum_{\mathbf{Q}\in\mathbf{Q}_w} , P(\mathbf{X}|\mathbf{Q})P(\mathbf{Q}|\mathbf{w}) \tag{1}$$

where $\mathbf{Q}$ is a sequence of phonemes and $\mathbf{Q}_w$ is the set of all possible state sequences that correspond to the word sequence $\mathbf{w}$, as defined by a lexicon (i.e. pronunciation dictionary) [9]. Then, $P(\mathbf{Q}|\mathbf{w})$ gives the probability of observing a certain phoneme sequence belonging to a word.

The speech recognition framework we use, Kaldi [10], constructs the decoding graph via Weighted Finite State Transducers (WFSTs) [11]. The final decoding graph $HCLG$ is a composition of multiple finite-state transducers:

$$HCLG = H \circ C \circ L \circ G, \tag{2}$$

where $\circ$ is the operation of graph composition for finite-state transducers (FST), and each element in Equation 2 represents a FST [11]. In summary, the phoneme posteriors are obtained via the acoustic model in H, and the HMM phone states are converted/relabeled to context-dependent 'triphone' states via C. Finally, the operation $L \circ G$ pairs any word string w in a pronunciation lexicon to its corresponding pronunciation q^w.

While it is common to use the standard CMUSphinx English Pronunciation Dictionary [12] in lyrics transcription [4], [6], a vowel-extended version has also been shown to be beneficial for word recognition [8]. In this paper, we aim to study the pronunciation variances in singing through a confusion analysis, a procedure similar to the ones presented in [13], [14]. Based on profound phonemic confusions, we create an extended version of the standard CMU dictionary by adding alternative word pronunciations for singing, which has been shown to be an effective method in ASR [15].

III. Pronunciation Analysis

The phonetic analysis is based on the confusions between orthographic transcriptions, $\widehat{\mathbf{Q}}$, produced by a pretrained ALT model that uses a pronunciation dictionary for speech (CMUSphinx [12]) and the human phoneme annotations, $\mathbf{Q}$, on the singing performances chosen for analysis. We use the *NUS Sung and Spoken Lyrics Corpus* [16] due to the availability of phoneme-level annotations, and choose native English speaking (with North American accent) singers *f01, f02, m09, m11* for analysis. We limit our analysis to these singers in order to minimize the influence of non-native accents.

Initially, the word transcriptions $\widehat{\mathbf{W}}$ are extracted and decomposed into their phonemic representations $\widehat{\mathbf{Q}}$ by decomposing the lexicon transducer L from the decoding graph $HCLG$. To get the phoneme confidences, we align $\widehat{\mathbf{Q}}$ with their corresponding manually annotated phoneme sequences. During this alignment, we take the following steps:

1) We compute the alignment score matrix $\mathbf{D}$ by performing Levenshtein alignment, *lev*, between the phoneme tokens q of the predictions $\widehat{\mathbf{Q}}_M$ and the ground truth $\mathbf{Q}_N$:
$$\mathbf{D}_{M\times N} = lev(\widehat{\mathbf{Q}}_M, \mathbf{Q}_N), \tag{3}$$
and find the best alignment path, $\mathbf{A}_{2\times K}$ through reverse tracing to find the path with the lowest pairwise gap cost:
$$\mathbf{A}_{2\times K} = \begin{pmatrix} \cdots & q_{k-1} & q_k & q_{k+1} & \cdots \\ \cdots & \widehat{q}_{k-1} & \widehat{q}_k & \widehat{q}_{k+1} & \cdots \end{pmatrix}. \tag{4}$$
$\mathbf{A}$ can be interpreted as a sequence of phoneme pairs.
2) There are three operations defined on these phoneme pairs to match $\widehat{\mathbf{Q}}_M$ to $\mathbf{Q}_N$: insertions (I), substitutions (S) and deletions (D). These operations are represented in $\mathbf{A}$ with the symbol ϵ. An alignment instance $a_k = \begin{pmatrix} \epsilon \\ \widehat{q}^*_k \end{pmatrix}$ is a deletion and the opposite case would be an insertion.
3) Let the number of correctly matching pairs in $\mathbf{A}$ be C, then the confidence score per phoneme type, c_q, can be retrieved as:
$$c_q = \frac{\sum_i^T C_{q,i} - (S_{q,i} + I_{q,i} + D_{q,i})}{\sum_i^T C_{q,i} + S_{q,i} + I_{q,i} + D_{q,i}}, \quad q \in \Omega_E, \tag{5}$$
where T is the number of utterances in the analysis set and Ω_E is the English phoneme set used in our analysis. The denominator is necessary to normalize with respect to the total number of pairs in $\mathbf{A}$, since the phonemes in Ω_E are not necessarily represented equally in the analysis data set.

Vowels	q	$c_q(R)$	Φ'_N
Short Vowels	AE	-0.42 (38)	AH, EH, AA
	AH	0.17 (33)	AA,EH,OW
	EH	0.3 (32)	AH,AE,IH
	IH	0.48 (26)	IY,AH,EY
	UH	0 (36)	AO,UW,AH
Long Vowels	AA	0.5 (24)	AO,AW,AE
	AO	0.06 (35)	AA,AH,OW
	ER	0.36 (31)	AH,OW,EH
	IY	0.87 (6)	EY,IH,EH
	UW	0.88 (4)	OW,AH,UH
Diphthongs	AY	0.86 (8)	AA,AH,EH
	AW	0.71 (18)	AA,AH
	EY	0.87 (7)	IY,AY,EH
	OW	0.76 (17)	AO,AA,AH
	OY	0.4 (28)	OW,AO,AY

TABLE I: Results of the phonetic analysis (vowels)

Tables I and II show the results of the phoneme confusion analysis for vowels and consonants respectively. The first two columns from the left are the list of English phoneme categories and types[1]. In the middle column, the confidence scores and their confidence rankings R are provided. By definition in Equation 5, $-1 \leq c_q \leq 1$, hence we did not further normalize this value. According to Equation 5, $c_q < 0.25$ means that there are less true positives than the sum of false negatives and positives in per phoneme type predictions, i.e. in most cases, q is predicted incorrectly. The phonemes in the rightmost column, Φ', are determined according to the most frequent instances of substitutions.

Consonants	q	$c_q(R)$	Φ'_N
Plosives	B	0.77 (16)	D,P,W
	D	0.16 (34)	T,N,JH
	G	0.77 (15)	NG,K
	K	0.85 (15)	G,HH
	P	0.78 (14)	B,M,F
	T	0.37 (29)	D,S,CH
Affricates	CH	0.79 (13)	JH,SH,T
	JH	0.88 (5)	CH,S,Y
Nasals	M	0.93 (2)	N,NG
	N	0.85 (12)	M,NG,D
	NG	0.85 (9)	N,M,T
Fricatives	DH	0.36 (30)	TH,D,N
	F	0.91 (3)	V,P,TH
	HH	0.70 (19)	DH,W,Y
	S	0.95 (1)	Z,TH,T
	SH	0.85 (10)	CH,S,Z
	TH	0.57 (21)	S,T,DH
	V	0.56 (22)	F,R,DH
	Z	-0.05 (37)	S,T
	ZH	N/A	N/A
Approximants*	L	0.44 (27)	AA,OW,AH
	R	0.48 (25)	AA,AH,IH
	W	0.66 (20)	AA,OW,V
	Y	0.55 (23)	IH, AH, IY

TABLE II: Results of the phonetic analysis (consonants)

It can be seen from the aforementioned tables that it is mostly the vowels that have the lowest confidences. Among vowels, *diphthongs* are more accurately predicted than *short vowels* and *long vowels*. The phoneme '*AE*' has the lowest c_q and is generally associated with the *schwa* sound in phonetics [17]. This very low c_q is not surprising as it is

[1]We use the standard 39-phoneme set of the CMU dictionary.

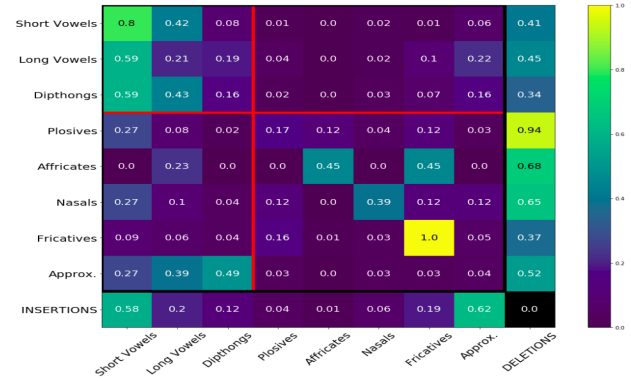

Fig. 1: Confusion matrix w.r.t. phoneme categories in Tables I and II. The red lines highlight the within-class regions for vowels and consonants. The numbers in cells are normalized values. The labels on the horizontal and the vertical axes represent the ground-truth and predictions respectively.

often pronounced weakly and is one of the most frequently occurring vowel sounds in the English language [18].

For consonants, c_q has higher values in general, but does not seem to be very consistent per phoneme category. The plosives '*D*' and '*T*' are severely confused indicating a systematic error, similarly for the other phonemes '*DH*', '*TH*' and '*Z*'. On the other hand, plosives '*B,G,K,P*' have rather high confidences. This is not extremely surprising as it has been mentioned in the literature that singers may utilize such phonemes to utter strong note offsets during melody construction [19]. According to this observation, the singers in our analysis data did not seem to omit '*B,G,K,P*' sounds. Though Φ' in different parent phoneme categories is not considered in Table II, we observed systematic confusions in *approximants* with vowels. This might be an indication of either systematic misalignment errors due to longer vowels, or omitted vowels for fluency during melody construction. In addition to the *short vowels*, the '*HH*' and '*Y*' sounds are inserted the most to the predictions compared to the manual human annotations.

In Figure 1, we show the phoneme confusion matrix summarized with respect to phonetic categories. We discard C_q for calculating the confusions and sum only S, I and D values for each phonetic category. Therefore the diagonal axis does not represent self-confidences. Instead it represents the domestic confusions within each phonetic category. Phonetic-category-wise normalization is applied based on unit sum. These normalization steps are crucial to get confusion values independent of the number of occurrences. Insertions and deletions for each category are also included in the figure. The concentration of high confusion rates can be observed for vowels (top left). Short vowels are mostly confused with *short vowels*. The annotated longer vowels are not necessarily represented in the standard speech lexicon, thus causing the system to assign a higher likelihood for the short vowels when making word predictions. Note the high number of deleted *plosives* signaling them being omitted from pronunciations during singing. Overall, a high frequency of deletions is observed. In addition to alignment errors, one possible cause for this could be the word liaisons being annotated as single phonemes in human annotations whereas the ALT system would predict such instances as separate phonemes. For example, in '*DREAM MAKER*', '*M*' is annotated once in the corresponding $\mathbf{Q}$, but detected twice in $\widehat{\mathbf{Q}}$.

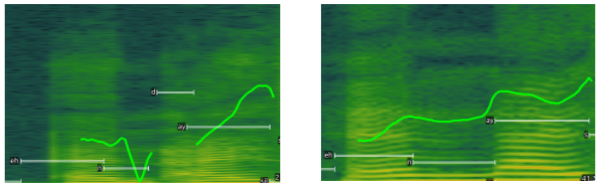

Fig. 2: An example of an omitted plosive in singing. W = 'AND I ' ; Q^{read} = 'AE N **D** AY' (left) ; Q^{sing} = 'EH N AY'. The gray horizontal lines show the temporal phoneme regions and the bright green curves are the pitch tracks extracted using pYIN [20].

IV. Extending the Lexicon

In this section, we propose a pronunciation model for sung utterances based on the observations of the previous step. We extend the standard pronunciation model for speech through generating alternative pronunciations for singing.

It is not seldom that in singing, performers may omit some consonants at the endings of words. This phenomenon can be explained as a stylistic convention that singers exhibit in their performances in order to maintain the sonority of their singing [1], or it could as well be a microphone technique to avoid unpleasant pops. The analysis in Section III suggests that this occurs most likely for *plosives* as the phoneme category with highest number of deletions. An example of an omitted *plosive* is illustrated in Figure 2[2]. The spectrogram segments in Figure 2 show the same words uttered as speech (left) and singing (right) by the same performer. According to the human annotators, the phoneme '*D*', is not present during singing. This can also be seen from the discrepancies in the spectrogram and the undisturbed pitch curve in the singing segment[3]. We add alternative pronunciations to such words ending with consonants *D, T, DH % Z* by removing their last phoneme in the corresponding q^{w_l}. We have chosen these consonants due to them having the lowest confidences according to the analysis in Section III .

It is noted in [3] that longer vowels in singing may potentially cause alignment errors, consequently affecting the training and thus the recognition performance. Gupta et al. [8] proposed to extend the occurrences of vowels in each word in the lexicon for modeling longer vowels. Through representing longer vowels as consecutive repeated phonemes

[2]The analysis is performed on Sonic Visualizer software [22].

[3]According to the empirical study in [21], pitch and phoneme perception are found to be cognitively correlated processes. Hence, we have chosen explicitly to show the pitch tracks.

a better alignment in singing performances can be achieved, and hence a potential improvement in WER. In this study, we also apply a similar strategy when extending the lexicon. For instance, consider the word *OCEANS* with its phonemic representation *OW SH AH N Z* in the lexicon. We extend the occurrence of each vowel for up to 2 times, for example: *OW OW SH AH N Z*, *OW SH AH AH N Z* instead of 4 times (as in [8]) so that a smaller transducer is generated, for efficient decoding.

The final version of the singing-adapted lexicon is constructed by combining the two approaches mentioned in this section. The goal of this is to create a model that is generalizable to common pronunciation variances observed in Section 3.

V. Experimental Setup

We evaluate the effectiveness of the proposed singing-adapted lexicon with respect to word recognition rate via ALT experiments. We compare its performance in terms of word and character error rates (WER) with a model trained on the standard CMU English pronunciation dictionary.

For training, we utilize the train split of the *DAMP* data set used by Demirel et al. [4] which consists of approximately 150 hours of monophonic singing recordings of English language pop songs recorded in a Karaoke setting with a non-negligible proportion of noise. There are performers from 30 different countries in the data set, hence allowing a powerful acoustic model to generalize the accentual variations.

	Char.	Words	Sentences	Recordings
NUS_read	21935	5788	781	32
NUS_sing	21935	5788	1029	32
DAMP_test	17609	4626	479	70

TABLE III: Statistics of evaluation sets

For testing the lexicons, we have trained the lyrics transcriber using the pipeline in [4] from the beginning at each experimental iteration. The transcriber in [4] is based on a hybrid-ASR framework where the acoustic model consists of neural networks trained on lattice-free maximum mutual information (LF-MMI) setting [23]. The neural network consists of stacks of 2D fully convolutional and factorized time-delay layers [24] with a self-attention layer added on top. At the input of the network, we extract 40-band filterbank features obtained with a hop size of 10ms and frame length of 20ms. To perform singer-adaptive training, we combine filterbank features with iVectors [25]. The phoneme posterior probabilities learned by the acoustic model are then decoded into a word-level representation with L and the grammar information (i.e the language model), G. We use a 4-gram language model (LM) using the SRILM toolkit [26] trained on the same lyrics corpus with the ones in [4], [6] which consists of recent English pop songs.

Results are reported on three evaluation sets (see Table III). The first set is the test split of the *DAMP - Sing! 300x30x2* data set [4] provided by Dabike et al. [6]. Other evaluation sets are the sung ("*NUS_sing*") and spoken ("*NUS_read*') splits of the NUS corpus excluding the native English speakers used in the phoneme analysis in Section 3. For experiments, we have manually segmented the NUS Corpus on the sentence level.

VI. Results

In the first stage of experiments, we test the benefit of different lexicon extension methods. In Table IV, L_{CMU} denote the standard CMU lexicon. L_1 and L_2 stand for extended lexicons where alternative pronunciations are generated via removing omitted (low-confidence) consonants and extending vowels (as explained in Section IV) separately. L_3 is then the final singing adapted lexicon which is a combination of both extension methods.

	L_{CMU}	L_1	L_2	L_3
DAMP_test	17.01	16.52	15.85	**15.49**
NUS_read	9.83	9.35	9.65	**9.40**
NUS_sing	11.57	10.61	10.30	**9.80**

TABLE IV: WERs of different lexicon variants

These initial results show that proposed lexicon extension methods are overall beneficial for sung word recognition, although L_1 resulted in rather more marginal improvement compared to L_2. Combining both extension methods achieved the best performance with a relative improvement of 8.24% WER w.r.t to L_{CMU}.

Further in Table V, we provide the word and character recognition results where the main comparison is between the recognition performances using the standard CMU dictionary and our singing-adapted version (L_3), in terms of the error (ER), substitution (S), insertion (I), deletion (D) rates explicitly. The singing-adapted dictionary performs consistently better than the speech dictionary even though it can be considered a modest improvement. Note that most of the improvements come from the reduced number of deletions, while the improvement in insertions is generally marginal.

	L_{CMU}				L_3			
	ER	*S*	*I*	*D*	*ER*	*S*	*I*	*D*
word								
DAMP_test	17.21	10.67	1.43	5.66	**15.49**	10.73	1.53	3.12
NUS_read	10.51	7.52	1.07	1.91	**9.40**	6.53	1.07	1.80
NUS_sing	13.19	8.60	1.63	2.95	**9.80**	6.90	1.26	1.54
character								
DAMP_test	11.41	4.78	1.85	4.79	**9.41**	4.25	1.63	3.53
NUS_read	5.57	2.73	1.38	1.47	**5.11**	2.54	1.11	1.36
NUS_sing	7.05	3.02	1.58	2.33	**6.14**	3.03	1.36	1.75

TABLE V: Word and character error rates using standard (*L_CMU*) and singing-adapted (*L_sing*) pronunciation dictionaries.

According to Table VI, L_3 shows more than absolute 5% lower ER on singing data. Less words are substituted and deleted using L_3. The vowel recognition rate is obtained via

[4]The data set is available for research upon request at https://ccrma.stanford.edu/damp.

	L_{CMU}			L_3		
	ER	*S*	*D*	*ER*	*S*	*D*
word (ending with consonants *D,DH,T,Z*)						
DAMP_test	22.84	13.06	7.78	**17.67**	10.15	7.21
NUS_read	9.74	8.82	0.91	**9.01**	7.90	1.10
NUS_sing	14.01	7.76	5.73	**7.94**	5.73	2.21
vowel						
DAMP_test	13.20	6.47	6.72	**9.80**	5.59	4.21
NUS_read	4.02	2.44	1.58	**3.99**	2.55	1.44
NUS_sing	7.23	2.98	4.26	**6.71**	3.03	3.68

TABLE VI: Error analysis w.r.t omitted consonants and vowels

comparing vowels in the human phoneme annotations and the phonetic transcript of the recognizer. The phonetic transcript is obtained similarly as explained in Section 3. The singing-adapted dictionary also performs consistently better than the speech version with regards to the vowel recognition rate although the improvement in the speech data is marginal, similarly for its response to words ending with low-confidence consonants. This shows that the adaptation is more singing specific, however the improvement is rather modest.

There are further possibilities for adapting the pronunciation model to singing. New alternative pronunciations may be generated via a statistical analysis of the interchange (substitution) of phonemes between speech and singing. Note that these substitutions need to be considered as context-dependent via observing the neighbouring phonemes for the instances of substitution. Additionally, pronunciation probabilities could be extracted from the training data which is reported to be beneficial for word recognition [27].

VII. Conclusion

This paper presents a computational approach for an in-depth analysis on the pronunciation differences between singing and speech. The proposed confusion analysis is utilized in identifying systematic pronunciation variances on the phoneme-level. We proposed a new singing-adapted version of the standard CMU dictionary by adding alternative word pronunciations based on the findings of our analysis. We report the best WER scores for ALT from monophonic recordings using an n-gram language model. The error analysis validates our approach being consistently beneficial for sung word recognition. We have publicly shared sentence-level manual annotations on the *NUS Sung and Spoken Lyrics Corpus* to be used as a new benchmark evaluation set for lyrics transcription in monophonic recordings.[5]

References

[1] Johan Sundberg and Thomas D. Rossing, "The science of the singing voice", Journal of the Acoustical Society of America, 1990.

[2] Lloyd A. Smith, and Brian L. Scott. "Increasing the intelligibility of sung vowels." The Journal of the Acoustical Society of America, 1980.

[3] Anna M Kruspe, "Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing.," in International Society for Music Information Retrieval Conference (ISMIR), 2016.

[4] Emir Demirel, Sven Ahlbäck, and Simon Dixon, "Automatic lyrics transcription with dilated convolutional networks with self-attention," in International Joint Conference on NeuralNetworks (IJCNN), 2020.

[5]The annotations can be retrieved from https://github.com/emirdemirel/ALTA/s5/data.

[5] Daniel Stoller, Simon Durand, and Sebastian Ewert, "End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[6] Gerardo Roa Dabike and Jon Barker, "Automatic lyrics transcription from karaoke vocal tracks: Resources and a baseline system", in Interspeech, 2019.

[7] Chitralekha Gupta, Emre Yılmaz, and Haizhou Li, "Automatic lyrics transcription in polyphonic music: Does background music help?", in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[8] Chitralekha Gupta, Haizhou Li, and Ye Wang, "Automatic pronunciation evaluation of singing", in Interspeech, 2018.

[9] Mark Gales and Steve Young "The application of Hidden Markov Models in speech recognition", Foundations and Trends in Signal Processing, Now Publishers Inc., 2008.

[10] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Bur-get, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, PetrMotlicek, Yanmin Qian, Petr Schwarz, et al.,"The Kaldi speech recognition toolkit", in Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, 2011.

[11] Mehryar Mohri, Fernando Pereira, and Michael Riley, "Speech recognition with weighted finite-state transducers", in Springer Handbook of Speech Processing, Springer, 2008.

[12] Robert Weide, "The CMU pronunciation dictionary, release0.6", 1998.

[13] Omar Caballero Morales and Stephen Cox. "Modelling confusion matrices to improve speech recognition accuracy, with an application to dysarthric speech", Eighth Annual Conference of the International Speech Communication Association. 2007.

[14] Emre Yilmaz and Joris Pelemans "Automatic assessment of children's reading with the FLaVoR decoding using a phone confusion model", Proceedings in Interspeech, 2014.

[15] Martine Adda-Deckker and Lori Lamel, "The use of lexica in automatic speech recognition" Lexicon Development for Speech and Language Processing. Springer, Dordrecht, 2000.

[16] Zhiyan Duan, Haotian Fang, Bo Li, Khe Chai Sim, and Ye 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. IEEE, 2013.

[17] Daniel Silverman, "Schwa" The Blackwell Companion to Phonology, 2011.

[18] Peter Roach, "British English: Received pronunciation" The Journal of the International Phonetic Association, 2004.

[19] William R. Bauer "Scat singing: a timbral and phonemic analysis", Current Musicology, 2002.

[20] Matthias Mauch and Simon Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions", in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014.

[21] Raymond L. Goldsworthy, "Correlations between pitch and phoneme perception in cochlear implant users and their normal hearing peers." Journal of the Association for Research in Otolaryngology 2015.

[22] Chris Cannam, Christian Landone, and Mark Sandler, "SonicVisualiser: An open source application for viewing, analysing,and annotating music audio files", in Association for Computing Machinery (ACM) Multimedia International Conference, 2010.

[23] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI", in Interspeech, 2016.

[24] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, HainanXu, Mahsa Yarmohammadi, and Sanjeev Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural net-works", in Interspeech, 2018.

[25] George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny, "Speaker adaptation of neural network acoustic models using i-vectors", in IEEE Workshop on Automatic Speech Recognition and Understanding, 2013.

[26] Andreas Stolcke, "SRILM - an extensible language modeling toolkit", in Seventh International Conference on Spoken Language Processing, 2002.

[27] Guoguo Chen, Hainan Xu, Minhua Wu, Daniel Povey, and Sanjeev Khudanpur, "Pronunciation and silence probability modeling for ASR", in Sixteenth Annual Conference of the International Speech Communication Association, 2015.