Emir Demirel

Topic: Automatic Lyrics Transcription and Alignment

Supervisors : Simon Dixon, Sven Ahlback

Queen Mary University of London

MIP-Frontiers

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

PART 1 :

Computational Pronunciation Analysis and Modelling of Sung Utterances

&

PART 2 :

Low Resource Audio-to-Lyrics Alignment

PART 1 :

Computational Pronunciation Analysis and Modelling of Sung Utterances

Contents

The word recognition rates of industry-level ASR systems can be higher.

- Introduction - Background Info - Pronunciation Analysis - Experiments - Conclusion & Future Work

The word recognition rates of industry-level ASR systems can be higher.

There are a number of factors for lower rates in singing performances:

Background music

Sound effects

The utterance of words

vs.

Synthesized Vocals

Simultaneous utterance

Training data

The word recognition rates of industry-level ASR systems can be higher.

There are a number of factors for lower rates in singing performances:

Background music

The utterance of words

vs.

Simultaneous utterance

Sound effects

Synthesized Vocals

Training data

- **Introduction** - Background Info - Pronunciation Analysis - Experiments - Conclusion & Future Work

The word recognition rates of industry-level ASR systems can be higher.

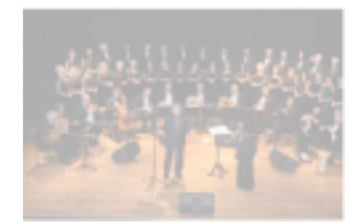
There are a number of factors for lower rates in singing performances:

Background music

The utterance of words

vs.

Simultaneous utterance



Training data


WELCOME TO THE DALI DATASET: a large Dataset of synchronised Audio, Lyrics and vocal notes.


DAMP-MVP: Digital Archive of Mobile Performances - Smule Multilingual Vocal Performance 300x30x2

Sound effects

Synthesized Vocals

Pronunciation in singing

- (Automatic) Lyrics Transcription is the process of recognising the most likely sequence of words from sung utterances.

- (Automatic) Lyrics Transcription is the process of recognising the most likely sequence of words from sung utterances.

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}(P(\mathbf{w}|\mathbf{X}))$$

Fundamental equation of (ASR)

- (Automatic) Lyrics Transcription is the process of recognising the most likely sequence of words from sung utterances.

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}(P(\mathbf{w}|\mathbf{X}))$$

OR applying the Bayes' rule

$$\boxed{\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}(P(\mathbf{X}|\mathbf{w})P(\mathbf{w})).}$$

- (Automatic) Lyrics Transcription is the process of recognising the most likely sequence of words from sung utterances.

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}(P(\mathbf{w}|\mathbf{X}))$$

OR applying the Bayes' rule

$$\boxed{\widehat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}(P(\mathbf{X}|\mathbf{w})P(\mathbf{w})).}$$

Acoustic Model

Language Model

- In Large Vocabulary Continuous Speech Recognition (LVCSR), the search space of words is very large.

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}(P(\mathbf{w}|\mathbf{X}))$$

- We need to take into account pronunciation variations

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}(P(\mathbf{X}|\mathbf{w})P(\mathbf{w})).$$

Thus, we build the acoustic model for **phonemes (Q)**;

$$\boxed{\widehat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}\, \underbrace{P(\mathbf{w})}_{\text{Language Model}} \sum_{\mathbf{Q}} \underbrace{P(\mathbf{X}|\mathbf{Q})}_{\text{Acoustic Model}} \underbrace{P(\mathbf{Q}|\mathbf{w})}_{\text{Pronunciation Model}}.}$$

$$\widehat{\mathbf{w}} = \operatorname*{argmax}_{\mathbf{w}} P(\mathbf{w}) \sum_{\mathbf{Q}} P(\mathbf{X}|\mathbf{Q}) P(\mathbf{Q}|\mathbf{w}).$$

Through Viterbi decoding, we find the most likely word sequence;

$$\widehat{\mathbf{w}} = \operatorname*{argmax}_{\mathbf{w}} P(\mathbf{w}) \max_{\mathbf{Q} \in Q_w} P(\mathbf{X}|\mathbf{Q}) P(\mathbf{Q}|\mathbf{w}),$$

Valid pronunciation of words defined by the lexicon (*)

$$P(\mathbf{Q}|\mathbf{w}) = \prod_{l=1}^{L} P(\mathbf{q}^{w_l}|w_l)$$

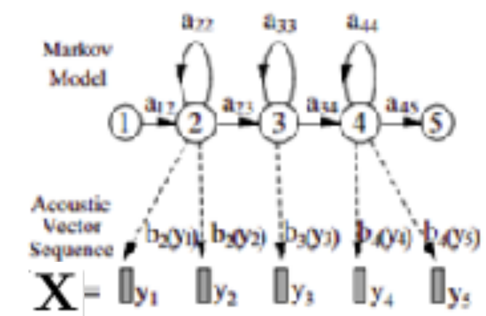
Phonetic Lexicon

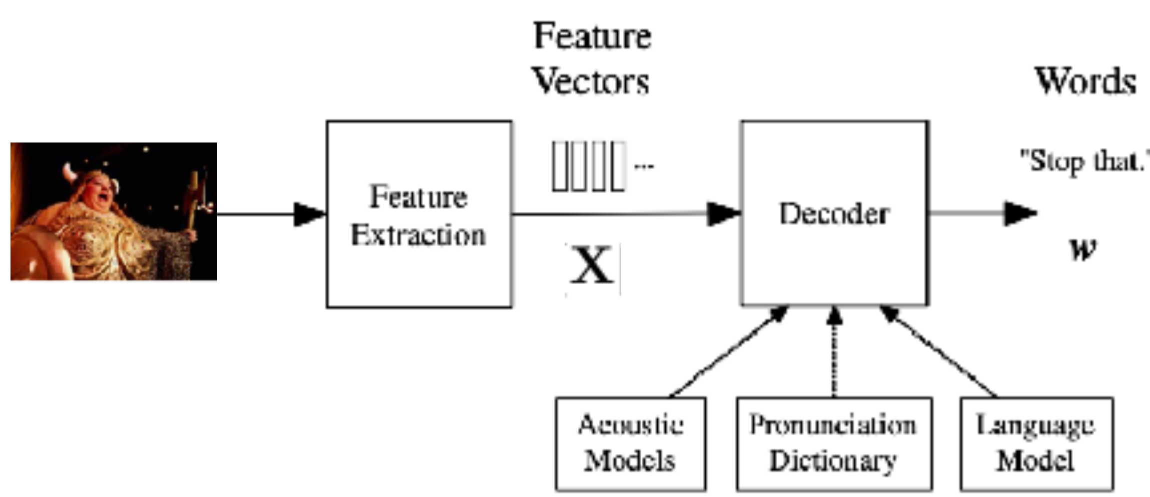
GIMME	G IH IH M IY
GIMME	G IH M IY IY
GIMME	G IH M IY
A	AA
MAN	M AE N
AFTER	AE F T ER
MIDNIGHT	M IH D N AY AY T

- Through vectorizing words with phoneme sequences, we allow multiple pronunciation variants of a word possible.

(*) Gales & Young (2008)

$$\widehat{\mathbf{w}} = \operatorname*{argmax}_{\mathbf{w}} P(\mathbf{w}) \max_{\mathbf{Q} \in Q_w} P(\mathbf{X}|\mathbf{Q}) P(\mathbf{Q}|\mathbf{w}),$$





1 Architecture of a HMM-based Recogniser.

Analysis Data: Native ENG speakers within the NUS Sung on Spoken Lyrics Corpus

Duan, Zhiyan, et al. "The NUS Sung and Spoken Lyrics Corpus: A Quantitative Comparison of Singing and Speech."

TABLE III
SUBJECTS IN THE NUS CORPUS

Code	Gender	Voice Type	Sung Accent	Spoken Accent
01	F	Soprano	North American	North American
02	F	Soprano	North American	North American
03	F	Soprano	North American	Mild Local Singaporean
04	F	Alto	Mild Malay	Mild Malay
05	F	Alto	Malay	Malay
06	F	Alto	Mild Malay	Mild Malay
07	M	Tenor	Mild Local Singaporean	Mild Local Singaporean
08	M	Tenor	Northern Chinese	Northern Chinese
09	M	Baritone	North American	North American
10	M	Baritone	North American	Standard Singaporean
11	M	Baritone	North American	North American
12	M	Bass	Local Singaporean	Local Singaporean

Example 4.1 *Substitutions in the utterance 'AND THE WONDER OF IT ALL'. w and $\widehat{w}$ are the human annotated ground truth and predicted word sequences. Q and $\widehat{Q}$ are the corresponding phonemic representations. In the bottom, the pronunciations obtained from the CMU English dictionary are provided. The word & pronunciation errors are highlighted with bold font.*

	w	AND	THE	WONDER	OF	IT	ALL
	$\widehat{w}$	AND	**THOUGH**	**ONE DARE**	**OUR**	**FEET**	ALL
Ground Truth	Q	HH EH N **eps**	D OW	W AH N D EH R	AO F	IH T	AO AH
Prediction	$\widehat{Q}$	**eps** AE N D	**DH** OW	W AH N D EH R	AA **R**	F IY T	AO **L**

We compute the alignment score matrix, $\mathbf{D}$, by performing Levenshtein alignment, lev between the phoneme sequences of the predictions, $\widehat{\mathbf{Q}}_M$ and the ground truth $\mathbf{Q}_N$,

$$\mathbf{D}_{M\times N} = lev(\widehat{\mathbf{Q}}_M, \mathbf{Q}_N) \tag{7}$$

and find the best alignment path, $\mathbf{A}_{2\times K}$ through reverse tracing to find the path with the lowest pairwise gap cost:

$$\mathbf{A}_{2\times K} = \begin{pmatrix} \cdots & \phi^*_{k-1} & \phi^*_k & \phi^*_{k+1} & \cdots \\ \cdots & \widehat{\phi^*}_{k-1} & \widehat{\phi^*}_k & \widehat{\phi^*}_{k+1} & \cdots \end{pmatrix}$$

Example 4.1 *Substitutions in the utterance 'AND THE WONDER OF IT ALL'. w and $\widehat{w}$ are the human annotated ground truth and predicted word sequences. Q and $\widehat{Q}$ are the corresponding phonemic representations. In the bottom, the pronunciations obtained from the CMU English dictionary are provided. The word & pronunciation errors are highlighted with bold font.*

w	AND	THE	WONDER	OF	IT	ALL
$\widehat{w}$	AND	**THOUGH**	**ONE DARE**	**OUR**	**FEET**	ALL
Q	HH EH N **eps**	D OW	W AH N D EH R	AO F	IH T	AO AH
$\widehat{Q}$	**eps** AE N D	**DH** OW	W AH N D EH R	AA **R**	F IY T	AO **L**

Example 4.1 *Substitutions in the utterance 'AND THE WONDER OF IT ALL'. w and $\widehat{w}$ are the human annotated ground truth and predicted word sequences. Q and $\widehat{Q}$ are the corresponding phonemic representations. In the bottom, the pronunciations obtained from the CMU English dictionary are provided. The word & pronunciation errors are highlighted with bold font.*

	w	*AND*	*THE*	*WONDER*	*OF*	*IT*	*ALL*
	$\widehat{w}$	*AND*	***THOUGH***	***ONE DARE***	***OUR***	***FEET***	*ALL*
Ground Truth	Q	*HH EH N **eps***	*D OW*	*W AH N D EH R*	*AO F*	*IH T*	*AO AH*
Prediction	$\widehat{Q}$	***eps** AE N D*	***DH** OW*	*W AH N D EH R*	*AA **R***	*F IY T*	*AO **L***

Insertion

Deletion

Substitution

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 $a_k = \begin{pmatrix} \epsilon \\ \widehat{\phi}^*_k \end{pmatrix}$ is a deletion and the opposite case would be an insertion.

Example 4.1 *Substitutions in the utterance 'AND THE WONDER OF IT ALL'. w and $\widehat{w}$ are the human annotated ground truth and predicted word sequences. Q and $\widehat{Q}$ are the corresponding phonemic representations. In the bottom, the pronunciations obtained from the CMU English dictionary are provided. The word & pronunciation errors are highlighted with bold font.*

	w	AND	THE	WONDER	OF	IT	ALL
	$\widehat{w}$	AND	**THOUGH**	**ONE DARE**	**OUR**	**FEET**	ALL
Ground Truth	Q	HH EH N **eps**	D OW	W AH N D EH R	AO F	IH T	AO AH
Prediction	$\widehat{Q}$	**eps** AE N D	**DH** OW	W AH N D EH R	AA **R**	F IY T	AO **L**

Let the number of correctly matching pairs in $\mathbf{A}$ be C, then the confidence score per phoneme type, c_ϕ, can be retrieved as:

$$c_\varphi = \frac{\sum_i^T C_{\phi,i} - (S_{\phi,i} + I_{\phi,i} + D_{\phi,i})}{\sum_i^T C_{\phi,i} + S_{\phi,i} + I_{\phi,i} + D_{\phi,i}}, \quad \phi \subset \Omega_E \tag{8}$$

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 dataset.

Vowels	φ	$c_\phi(R)$	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 1: Results of the phonetic analysis (vowels)

Consonants	ϕ	$c_\phi(R)$	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 2: Results of the phonetic analysis (consonants)

$$-1 \leq c_\phi \leq 1$$

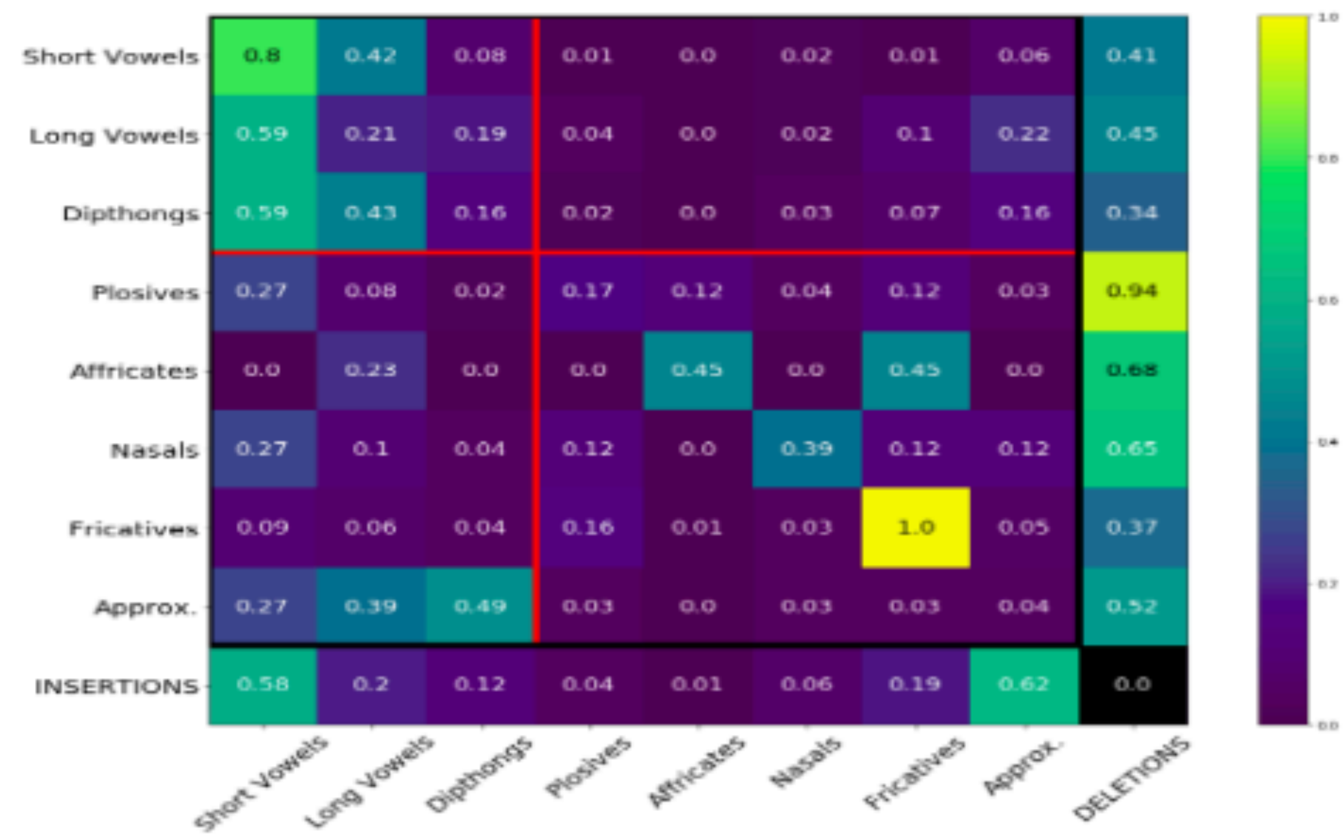


Fig. 1: Confusion matrix w.r.t. phoneme categories in Tables 1 and 2. The red lines highlight the within-class regions for vowels and consonants. The numbers in cells are normalized values. The labels on the horizontal and the vertical axes represent the ground-truth and predictions respectively.

- Introduction - Background Info - Pronunciation Analysis - Experiments - Conclusion & Future Work

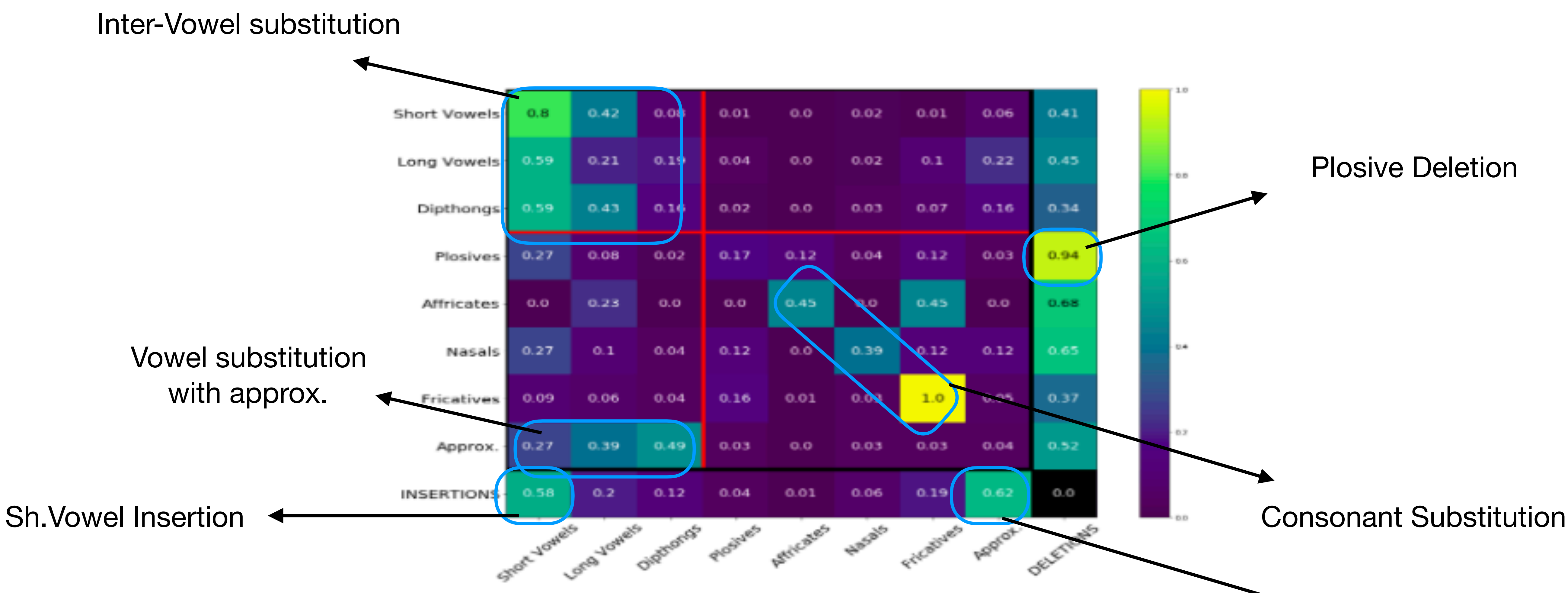


Fig. 1: Confusion matrix w.r.t. phoneme categories in Tables 1 and 2. The red lines highlight the within-class regions for vowels and consonants. The numbers in cells are normalized values. The labels on the horizontal and the vertical axes represent the ground-truth and predictions respectively.

Three observations when uttering words in singing:

- Longer vowels
- Omitted plosives & approximants
- Triphone substitutions

Three observations when uttering words in singing:

- **Longer vowels**
- Omitted plosives
- Triphone substitutions

	Speech	Singing
Duration (min)	0.03	0.18
Duration (max)	0.77	3.86
Duration (avg.)	0.10	0.34
Articulation Rate (per min)	266.25	172.50

Table 4: Mean articulations rates (syllables per minute) and duration stats (in seconds)

Figure 2: Duration distributions of vowels

Three observations when uttering words in singing:

- Longer vowels
- **Omitted plosives**
- Triphone substitutions

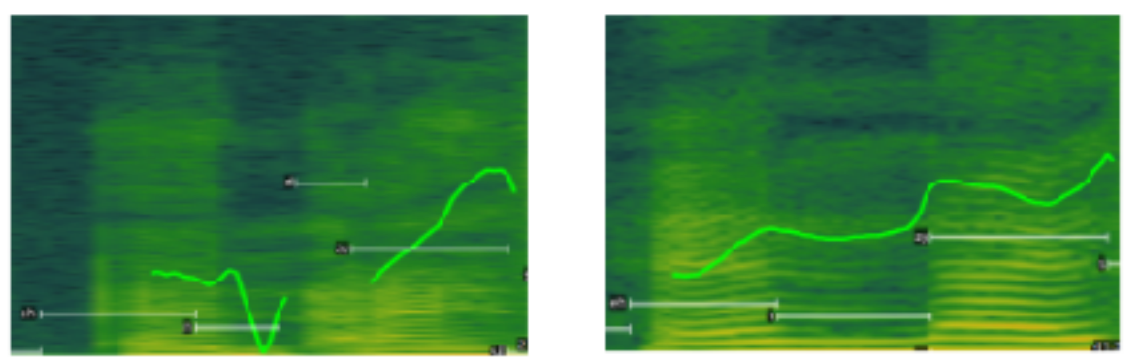


Fig. 2: An example of an omitted plosive in singing. W = 'AND I ' ; Q^{read} = 'ΛE N **D** ΛY' (left) ; Q^{sing} = 'EH N ΛY'. The gray horizontal lines show the temporal phoneme regions and the bright green curves are the pitch tracks extracted using pYIN [16].

Three observations when uttering words in singing:

- Longer vowels
- Omitted plosives
- **Triphone substitutions**

Vowels	φ	$c_\varphi(R)$	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 1: Results of the phonetic analysis (vowels)

Consonants	ϕ	$c_\phi(R)$	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 2: Results of the phonetic analysis (consonants)

We extend the pronunciations CMU dictionary with longer vowels and omitted plosives…

FLOAT	F	L	OW		T
FLOAT	F	L	OW	OW	T
FLOAT	F	L	OW		
FLOAT	F	L	OW	OW	T

FLOATING	F	L	OW		T	IH	NG	
FLOATING	F	L	OW	OW	T		IH	NG
FLOATING	F	L	OW		T	IH	IH	NG
FLOATING	F	L	OW		T	IH	N	
FLOATING	F	L	OW	OW	T		IH	N
FLOATING	F	L	OW		T	IH	IH	N

And test the effectiveness of these pronunciation extensions in the context of word recognition.

In the experiments, the s.o.t.a lyrics transcription framework is employed (Demirel, 2020) - open source toolkit!.

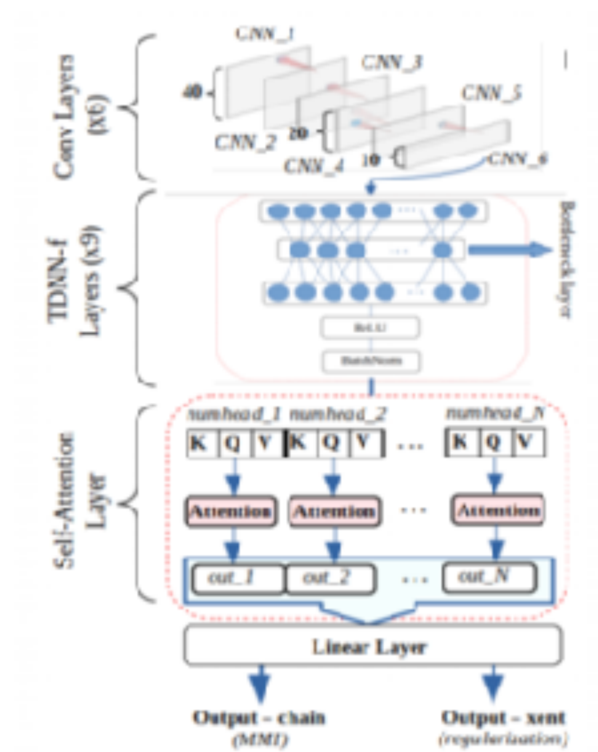
ALTA - (A)utomatic (L)yrics (T)ranscription & (A)lignment

A kaldi recipe for automatic lyrics transcription and audio-to-lyrics alignment tasks.

If you use this repository, please cite it as follows:

```
@inproceedings{demirel2020,
  title={Automatic lyrics transcription using dilated convolutional neural networks with self-attent:
  author={Demirel, Emir and Ahlback, Sven and Dixon, Simon},
  booktitle={International Joint Conference on Neural Networks},
  publisher={IEEE},
  year={2020}
}
```

Link to paper : https://arxiv.org/abs/2007.06486



Evaluation sets:

- A capella recordings
- Both sung and read
- Gender-wise balanced

Mixed (predominantly non-native)

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

Native from UK

Table 3: Statistics of evaluation sets

Speech lexicon

Singing adapted lexicon

	L_CMU				*L_sing*			
	ER	*S*	*I*	*D*	*ER*	*S*	*I*	*D*
word								
DAMP_test	17.21	10.67	1.43	5.66	**15.61**	10.83	1.53	3.24
NUS_read	10.51	7.52	1.07	1.91	**9.26**	6.41	1.01	1.80
NUS_sing	13.19	8.60	1.63	2.95	**10.06**	7.16	1.25	1.65
character								
DAMP_test	11.41	4.78	1.55	4.79	**9.88**	4.55	1.63	3.50
NUS_read	5.57	2.73	1.38	1.47	**5.10**	2.62	1.11	1.37
NUS_sing	7.07	3.13	1.60	2.33	**6.14**	3.03	1.36	1.75

Table 4: Word and character error rates using standard (*L_CMU*) and singing-adapted (*L_sing*) pronunciation dictionaries.

	L_CMU			*L_sing*		
	ER	*S*	*D*	*ER*	*S*	*D*
word (ending with plosives)						
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 5: Error analysis for plosives and vowels

Speech lexicon

Singing adapted lexicon

	L_CMU				L_sing			
	ER	S	I	D	ER	S	I	D
word								
DAMP_test	17.21	10.67	1.43	5.66	**15.61**	10.83	1.53	3.24
NUS_read	10.51	7.52	1.07	1.91	**9.26**	6.41	1.01	1.80
NUS_sing	13.19	8.60	1.63	2.95	**10.06**	7.16	1.25	1.65
character								
DAMP_test	11.41	4.78	1.55	4.79	**9.88**	4.55	1.63	3.50
NUS_read	5.57	2.73	1.38	1.47	**5.10**	2.62	1.11	1.37
NUS_sing	7.07	3.13	1.60	2.33	**6.14**	3.03	1.36	1.75

Table 4: Word and character error rates using standard (*L_CMU*) and singing-adapted (*L_sing*) pronunciation dictionaries.

	L_CMU			L_sing		
	ER	S	D	ER	S	D
word (ending with plosives)						
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 5: Error analysis for plosives and vowels

- Consistent WER improvement
- Marginal improvement for *read* samples

- Conclusion & Future Work

- A number of frequent pronunciation variances during singing are identified using a novel computational method that combines human annotations and an AI-based lyrics transcription system.

- Using a singing-adapted lexicon can yield to improvement in word recognition rates.

- Sentence-level annotations are provided for NUS Corpus, which can be leveraged for both training and evaluation in the context of automatic lyrics transcription.

- Conclusion & Future Work

- Add new pronunciations based on ‘triphone substitutions’.

- Obtain pronunciation probabilities.

Emir Demirel

Topic: Automatic Lyrics Transcription and Alignment

Supervisor: Simon Dixon

Queen Mary University of London

e.demirel@qmul.ac.uk

Q&A?