Computational Pronunciation Analysis and Modeling in Singing

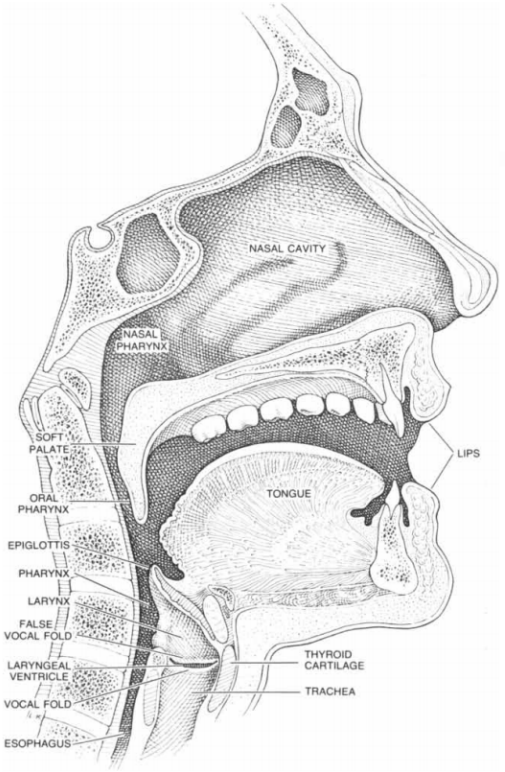
Emir Demirel, Sven Ahlback, Simon Dixon

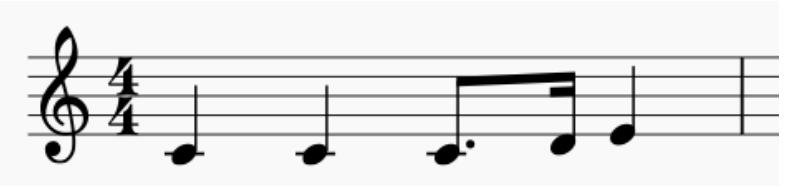
Contents

Introduction

- The performance of Automatic Lyrics Transcripton (singing) systems are far behind Speech Recognition (speech).
- Adapting ASR framework for singing
- Create a pronunciation dictionary for singing

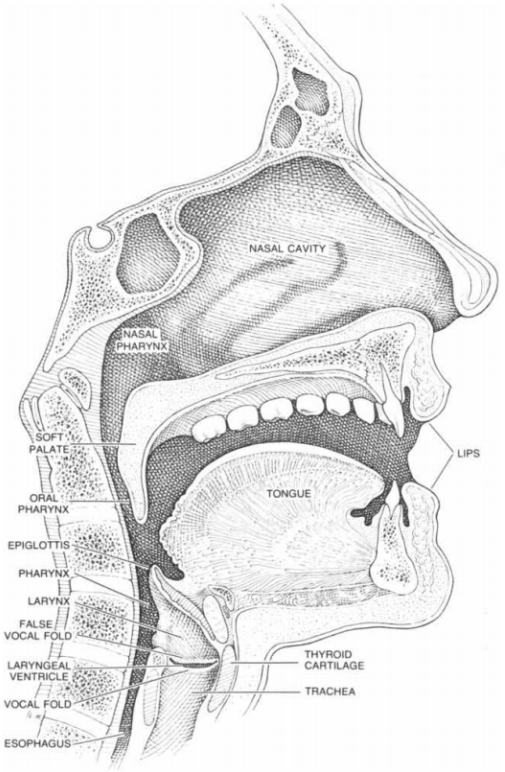
Introduction

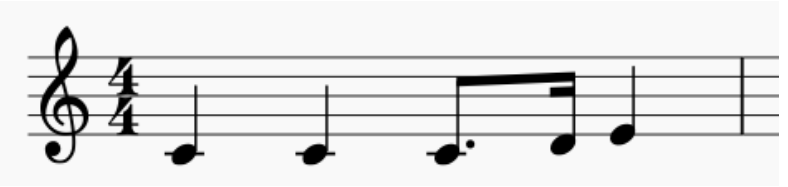




Vowels

Introduction





Row Row Row your boat

Background – Automatic lyrics transcription

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

or applying the Bayes Rule and with $P(\mathbf{X}) = 1$,

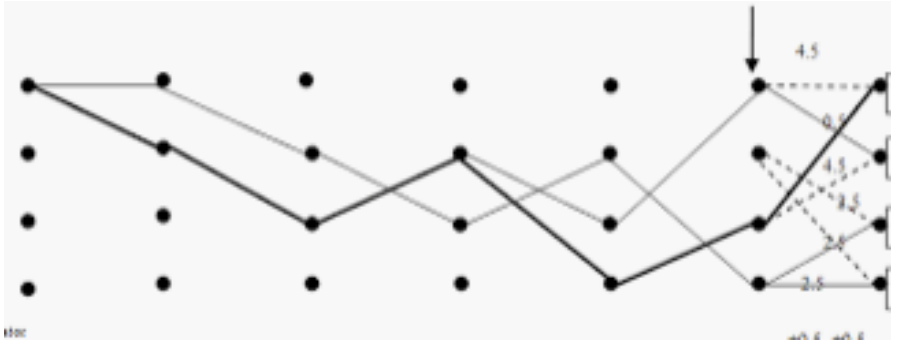
$$\widehat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}(P(\mathbf{X}|\mathbf{w})P(\mathbf{w}))$$

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

Background – Automatic lyrics transcription

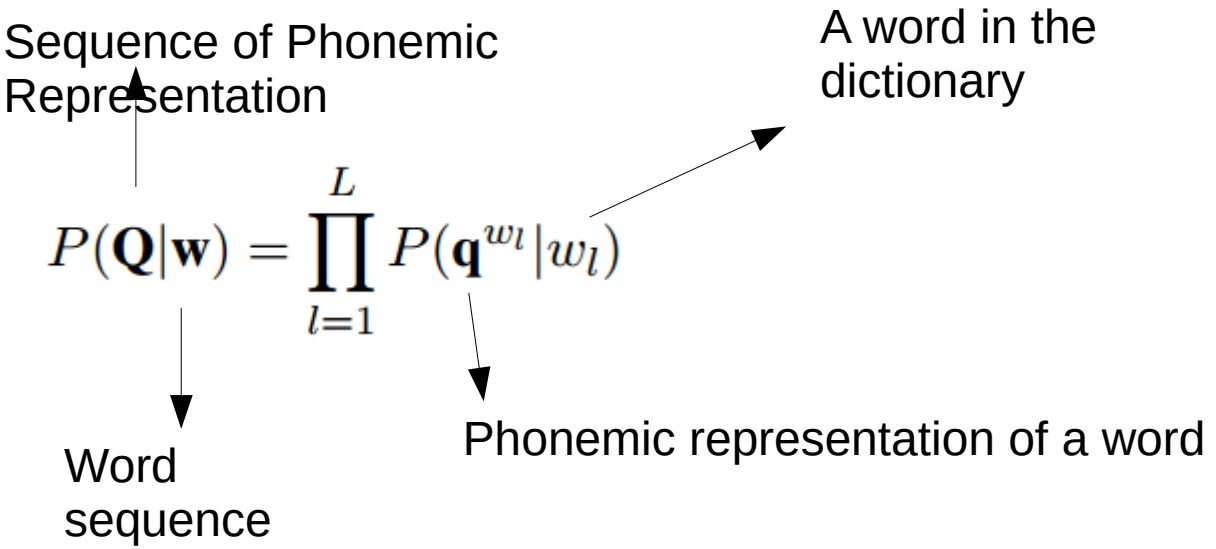
$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}\, P(\mathbf{w}) \sum_{\mathbf{Q}} P(\mathbf{X}|\mathbf{Q}) P(\mathbf{Q}|\mathbf{w})$$

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}\, P(\mathbf{w}) \max_{\mathbf{Q} \in Q_w} P(\mathbf{X}|\mathbf{Q}) P(\mathbf{Q}|\mathbf{w})$$



viterbi

Background – Automatic lyrics transcription



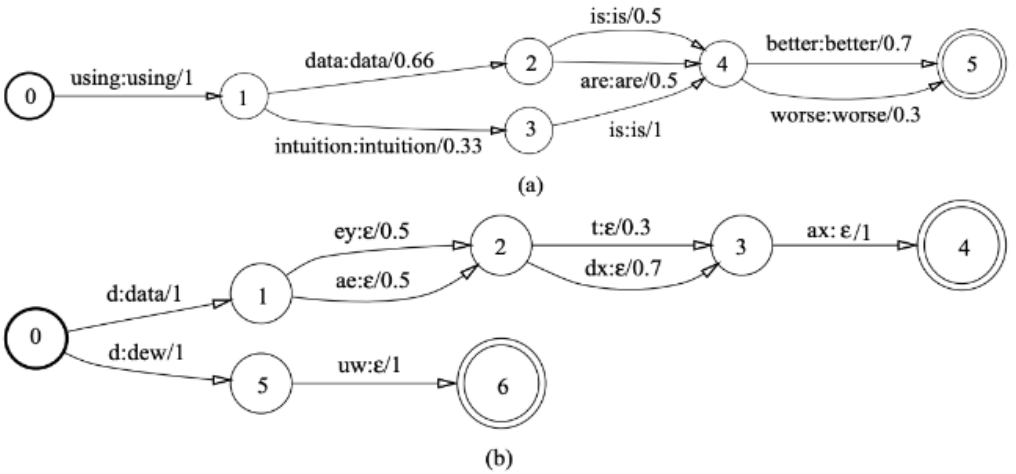
“

Background – Automatic lyrics transcription

Transducer	Function	Input sequence	Output sequence
H	acoustic model	HMM topology	context-dependent phonemes
C	context-dependency	context-dependent phonemes	phonemes
L	pronunciation dictionary	phonemes	words
G	language model	words	words

Table 1: Finite state transducers in the decoding graph $HCLG$

$$HCLG = min(det(H \circ C \circ L \circ G)) \tag{6}$$



Analysis

Based on ortographic transcriptions W(hat)

1. Each phoneme, ϕ, is converted to a single character, ϕ^* for robustness in text alignment,
2. 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():

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

$\mathbf{A}$ can be interpretted as a sequence of phoneme pairs.

Analysis

Based on ortographic transcriptions W(hat)

3. 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.

4. 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:

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

$$\phi \in \Omega_E \tag{9}$$

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.

Analysis

Vowels	ϕ	$c_\phi(R)$	Φ'_N
Short Vowels	AE	-0.70 (38)	AH, EH, AA
	AH	0.10 (32)	AA,EH,OW
	EH	0.34 (25)	AH,AE,IH
	IH	0.28 (28)	IY,AH,EY
	UH	-0.25 (36)	AO,UW,AH
Long Vowels	AA	0.36 (24)	AO,AW,AE
	AO	0.02 (34)	AA,AH,OW
	ER	0.1 (33)	AH,OW,EH
	IY	0.69 (17)	EY,IH,EH
	UW	0.78 (6)	OW,AH,UH
Diphthongs	AY	0.72 (13)	AA,AH,EH
	AW	0.57 (18)	AA,AH,*W
	EY	0.80 (5)	IY,AY,EH
	OW	0.70 (16)	AO,AA,AH
	OY	0.43 (22)	OW,AO,AY

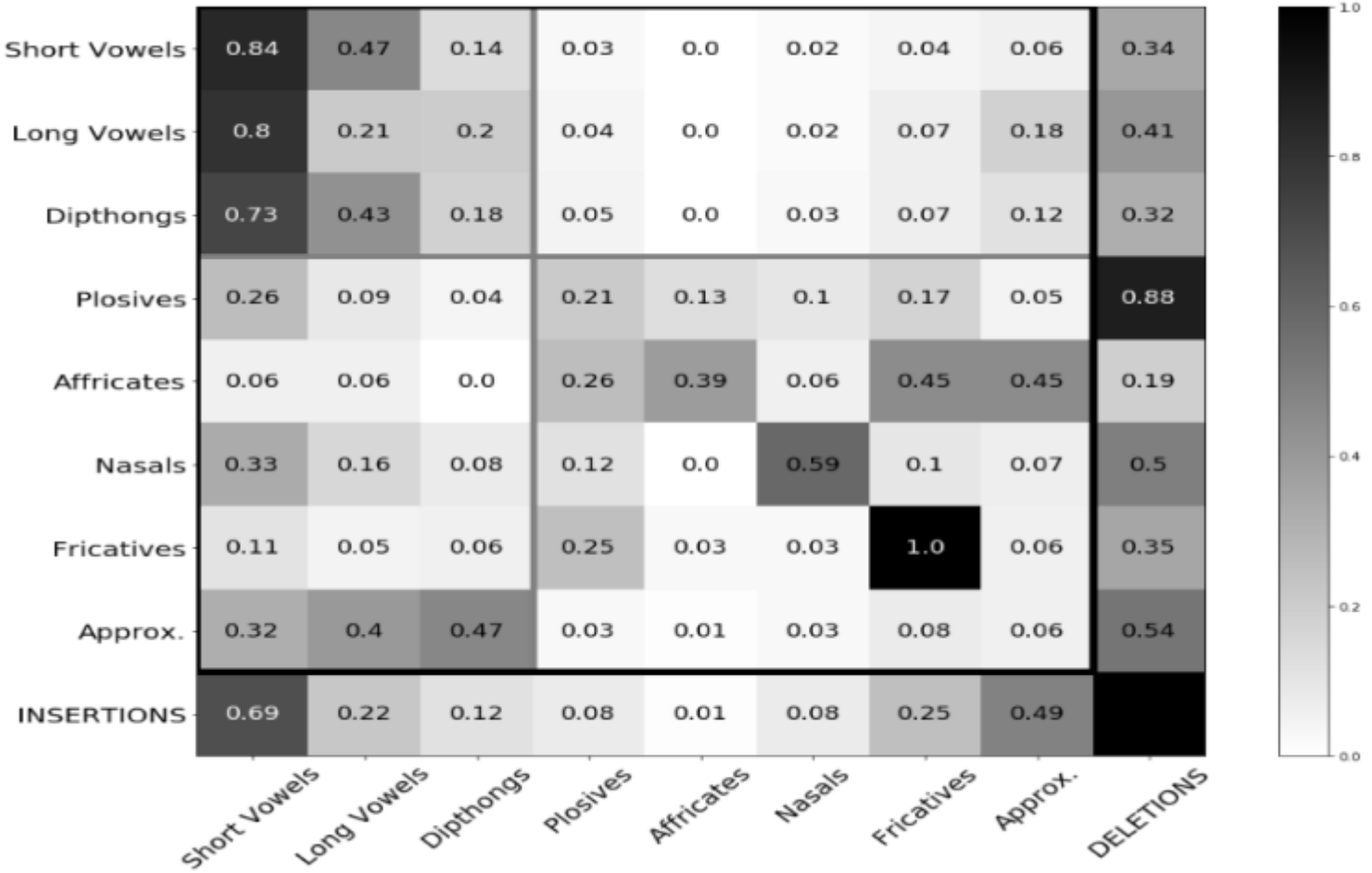
Table 2: The results of the phonetic analysis (Vowels)

Consonants	ϕ	$c_\phi(R)$	Φ'_N
Plosives	B	0.57 (19)	D,P,W
	D	-0.02 (35)	T,N,JH
	G	0.74 (11)	NG,K
	K	0.71 (15)	G,HH
	P	0.72 (14)	B,M,F
	T	0.2 (30)	D,S,CH
Affricates	CH	0.73 (12)	JH,SH,T
	JH	0.78 (7)	CH,S,Y
Nasals	M	0.86 (2)	N,NG
	N	0.75 (9)	M,NG,D
	NG	0.76 (8)	N,M,T
Fricatives	DH	0.16 (31)	TH,D,N
	F	0.81 (4)	V,P,TH
	HH	0.45 (21)	DH,W,Y
	S	0.89 (1)	Z,TH,T
	SH	0.81 (3)	CH,S,Z
	TH	0.23 (29)	S,T,DH
	V	0.39 (23)	F,R,DH
	Z	-0.37 (37)	S,T
	ZH	N/A	N/A
Approximants*	L	0.32 (26)	AA,OW,AH
	R	0.29 (27)	AA,AH,IH
	W	0.73 (10)	AA,OW,V
	Y	0.48 (20)	IH, AH, IY

Table 3: The results of the phonetic analysis (Consonants)

Analysis

Per Phoneme Category:



Analysis

Prosodic Features

	Speech	Singing
Articulation Rate (per min)	266.25	172.5
Duration (min)	0.03	0.18
Duration (max)	0.77	3.86
Duration (avg.)	0.1	0.34

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

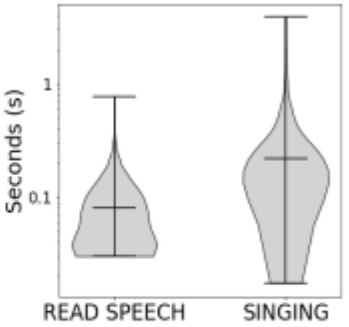


Figure 2: Duration distributions of vowels

Extending the Lexicon

Deletions

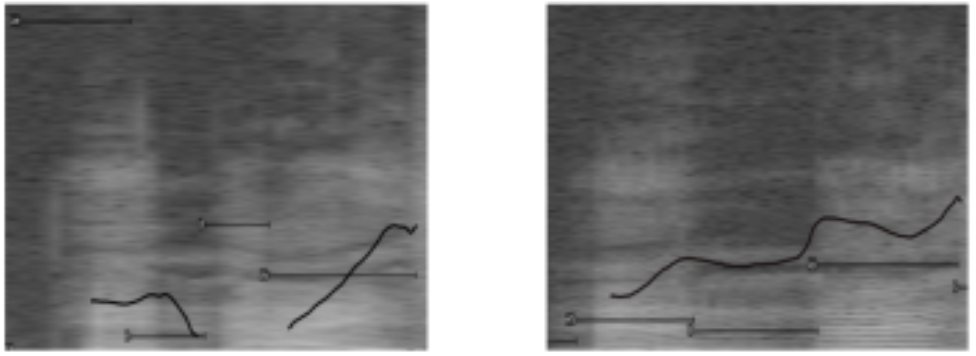


Figure 3: An example of an omitted plosive in singing. W = 'AND I ' ; Q^{read} = 'AE N **D** AY' (left) ; Q^{sing} = 'EH N AY'. The gray horizontal lines show the temporal phoneme regions and the bright green curves are the pitch track extracted using pYIN(Mauch and Dixon, 2014).

Extending the Lexicon

Insertions

OCEANS

OW SH AH N Z

OW OW SH AH N Z,
OW OW OW SH AH N Z,
OW OW OW OW SH AH N Z,
OW SH AH AH N Z

Extending the Lexicon

Substitutions

AY AY ; TH T ; AO AO

To create new pronunciations with substituted phonemes, we employ the following logic: First, for every substitution (ϕ_k, ϕ'_k) with $c_\phi < 0.8$ where $\phi' \in \Phi'$, we extract triphone patterns $\Xi(\phi)$ by concatenating the neighbouring phonemes of ϕ in $\mathbf{A}$:

$$\Xi(\phi) = (\phi_{k-1} \;\; \phi_k \;\; \phi_{k+1}) \tag{10}$$

and $\Xi(\phi')$:

$$\Xi(\phi') = (\phi_{k-1} \;\; \phi'_k \;\; \phi_{k+1}) \tag{11}$$

Then, we replace $\Xi(\phi)$ with $\Xi(\phi')$ in q^{w_L} to obtain the new alternate pronunciation q'^{w_l}.

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

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

Extending the Lexicon

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**
q^{w_L}	AE N D AH N D	DH AH DH IY	W AA N D ER W AH N D ER	AH V	IH T	AO L
$\widehat{q}^{w_L}$	-	DH OW	W AH N / D EH R W AA N / -	AA R AW ER	F IY T	-

ALT Experiments

TEST DATA :

DAMP – Sing! 300X30x2

NUS Sung and Spoken Lyrics Corpus – subsets:

- Read speech vs. singing
- Native vs. non_native

	L_0	L_1	L_2	L_3
NUS_READ	9.83	9.55	9.65	9.26
NUS_SING	11.57	10.94	10.30	10.89
NUS_NATIVE	7.97	7.43	7.08	7.20
NUS_NON-NATIVE	12.17	11.78	11.48	11.34
DAMP_TEST	17.01	16.71	16.26	16.73

Table 5: Lyrics Transcription Results (WER %)

Thanks