Automatic Pronunciation Evaluation of Singing

Chitralekha Gupta[1,2], Haizhou Li[3], and Ye Wang[1,2]

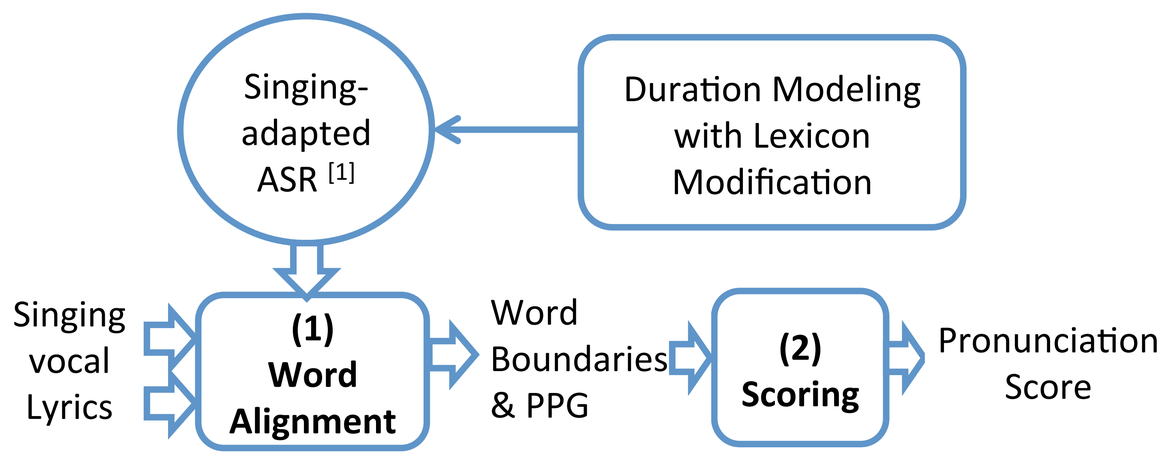
chitralekha@u.nus.edu, haizhou.li@nus.edu.sg, wangye@comp.nus.edu.sg

[1]School of Computing, [2]NUS Graduate School for Integrative Sciences and Engineering, [3]Department of Electrical and Computer Engineering, National University of Singapore

1. Motivation and Goal

- Pronunciation of lyrics is an important component of **singing performance**
- Singing is shown to improve pronunciation in **foreign language learning**
- Music and speech therapists apply **Melodic Intonation Therapy** to treat patients with speech disorders, such as non-fluent aphasia
- ***Goal:*** **To design a strategy to automatically evaluate pronunciation in singing voice**

2. A two-stage approach



3. Duration Modeling with Lexicon Modification

- Longer duration of vowels can be viewed as a type of *pronunciation variation*
- We modify the lexicon to model longer duration
- Eg.: the word “sleep” will have the lexicon variants:
  [S L IY P],
  [S L IY IY P],
  [S L IY IY IY P],
  [S L IY IY IY IY P]
- The ASR selects the closest matching variant at the time of forced-alignment

4. How does lexicon modification help?

Improves Word Alignment: 8% improvement from 83.7% to **91.7%** within 50 ms deviation (896 words across 100 utterances DAMP Karaoke dataset)

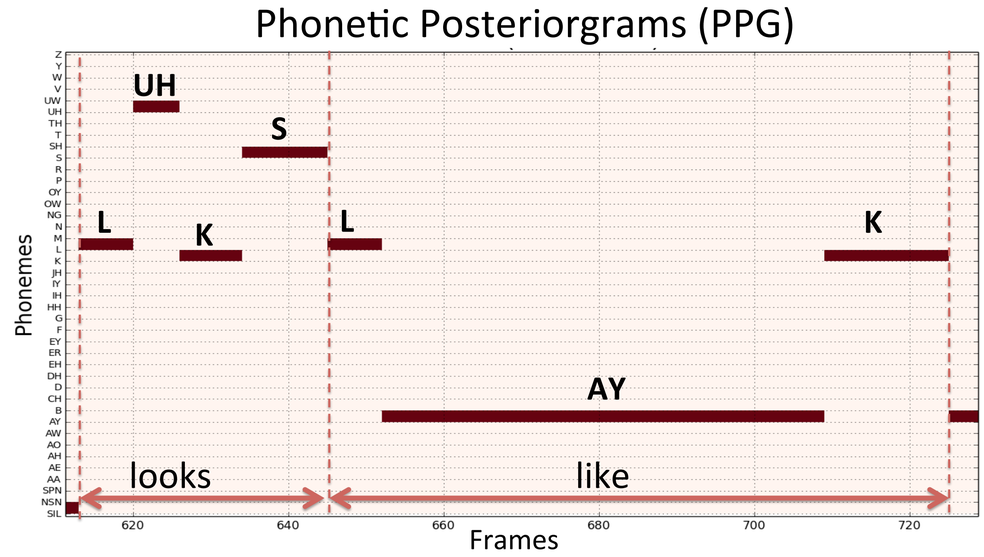
Singing-adapted models	<20ms	20-50ms	50-100ms	100-200ms	>200ms
w/o LEX modification	635	115	82	24	40
LEX modification	**748**	**74**	33	9	32

Improves Lyrics Recognition in singing voice

Models Adapted by Singing Data	%WER
w/o LEX modification	36.32
LEX modification	**29.65**

5. Scoring

Phonetic Posteriorgrams (PPG)



(1) Template independent PEM score indicates how close the pronunciation of a test sung utterance is to the **target lyrics**

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

(2) Template dependent PEM score indicates how close the pronunciation of a test sung utterance is to a **reference sung utterance**

$$PEM_{dep} = -\log(P_r \cdot P_t)$$

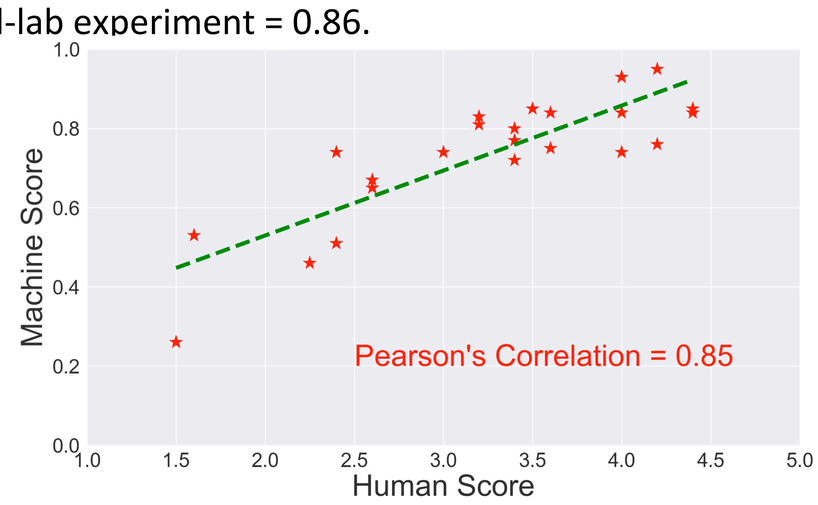
6. Word-level Scoring Validation Experiment

Data: 990 words across 100 utterances: 10 sung utterances from 10 singers, (5 from EN zone, and 5 from non-EN zone) (DAMP)
Ground-truth: binary pronunciation judgment per word by two university students fluent in English

Method: Template-	Accuracy	FPR	FNR
Dependent	0.52	0.48	0.47
Independent (LEX)	**0.72**	**0.28**	**0.28**
Independent (w/o LEX)	0.70	0.30	0.30

7. Song-level Scoring Validation Experiment

Data: 24 singers (DAMP) (13 female, 11 male) each singing one of 6 unique English popular songs
Ground-truth: Average song-level pronunciation judgments (over 5 raters) from crowd-sourcing platform Mturk. Pearson’s correlation with controlled-lab experiment = 0.86.



8. Contributions

- A strategy to compute reliable pronunciation evaluation scores for singing voice
- Duration-based lexicon modification for improvement in word alignment as well as scoring accuracy
- The pronunciation annotations dataset: https://drive.google.com/open?id=19JPEWSBAM0ssatjBIJzAzjClxi2abt8w

[1] Gupta, Chitralekha, Rong Tong, Haizhou Li, and Ye Wang. “Semi-supervised lyrics and solo-singing alignment”, ISMIR 2018.

Sound & Music Computing Lab
http://www.smcnus.org/

NUS National University of Singapore | School of Computing

NUS Graduate School for Integrative Sciences and Engineering
National University of Singapore