Towards Automatic Mispronunciation Detection in Singing

Chitralekha Gupta[1,2], David Grunberg[3], Preeti Rao[4], and Ye Wang[1]

chitralekha@u.nus.edu, david_grunberg@sutd.edu.sg, prao@ee.iitb.ac.in, wangye@comp.nus.edu.sg

[1]School of Computing, National University of Singapore, [2]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, [3]Singapore University of Technology and Design, [4]Indian Institute of Technology Bombay

1. Introduction

- **Learning a second language (L2) through singing** is shown to be effective and is used in pedagogy.
- Automatic pronunciation evaluation of singing is desirable for L2 learning, but finding training data is challenging.
- We propose a knowledge-based approach with limited data in an automatic speech recognition (ASR) framework to detect mispronunciation in singing.

I'm sitting here in the boring room
It's just another rainy Sunday afternoon
I'm wasting my time
I got nothing to do
I'm hanging around
I'm waiting for you
But nothing ever happens and I wonder

2. Problem statement

Pronunciation error detection in South-East Asian English accents singing (Malaysian: M, Indonesian: I, Singaporean: S) :

- What are the error patterns observed in non-native singing compared to non-native speech?
- If only native English speech trained phone models are available, can we detect pronunciation errors in non-native singing, given that we know the singer's L1 (native language)?

3. Error Patterns in Non-Native Singing

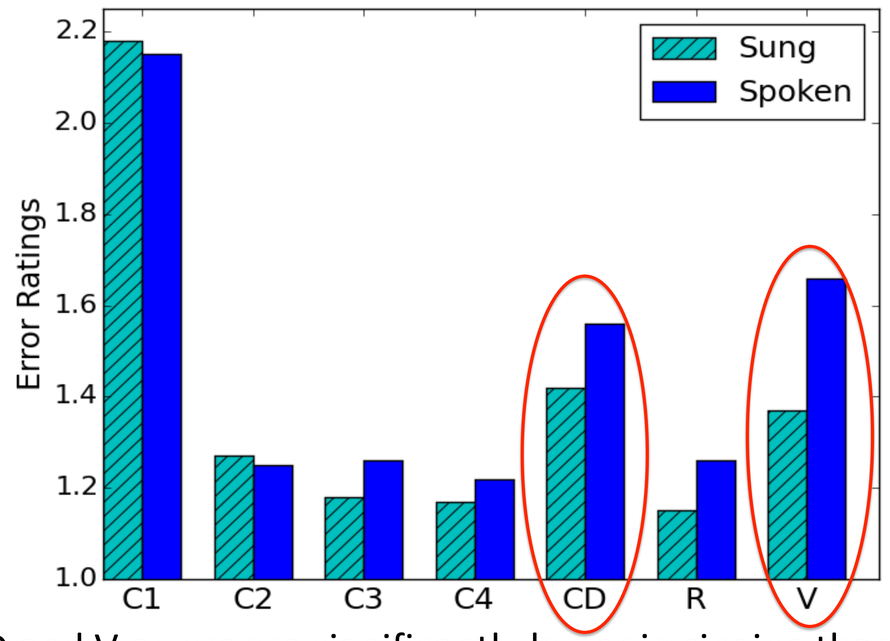
The typical error patterns reported in SE Asian English speech are as follows:

ID	Error	Examples
C1	/dh/→/d/	thy → die; mother → moder
C2	/th/→ /t/	thought → taught; nothing → noting
C3	/t/→ /th/	to→thu; sitting → sithing
C4	/d/→ /dh/	dear → dhear
CD	Word-end consonant deletion	moment → momen
R	Rolling /r/	ray → rray
V	vowel error	fool→full; sleeping→slipping

- **Are all of these error patterns also observed in singing?**

Dataset

- 26 sung and 26 spoken songs by 8 unique subjects (4M, 4F) - 3 Indonesian, 3 Singaporean, and 2 Malaysian.
- All of the above error patterns were subjectively rated by 3 judges: two native English speakers, one proficient in English – inter-judge agreement was high.

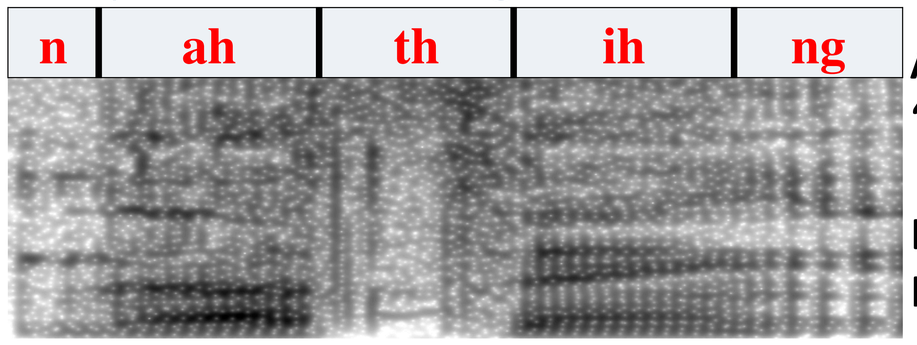


- CD and V errors are significantly lower in singing than in speech.
- Only a subset of the error patterns that occur in speech occur in singing.

This key insight suggests a possible learning strategy: learning this *subset* of phoneme pronunciation through singing, and the rest through speech.

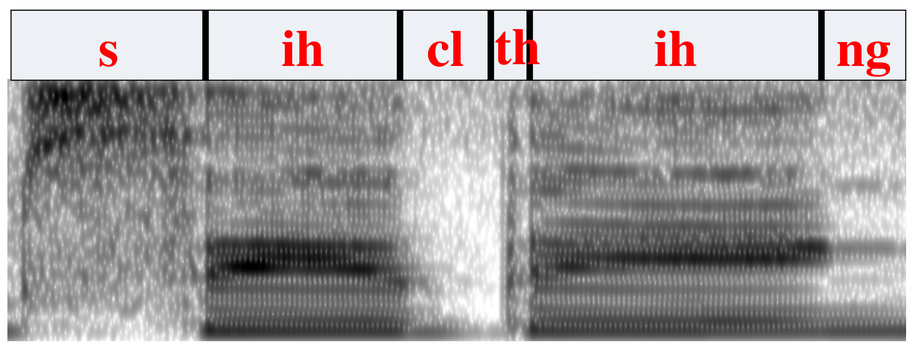
4. Mispronunciation Detection

Sub-phonetic Modeling



American "Nothing"

Dental Fricative /th/



Indonesian "Sitting" → "sithing"

Dental Stop /closure/+/th/

	Dictionary A	Dictionary B
Definition	only American English phones (L2)	American phones +modified (L1-adapted) phone
Example	/th/	/th/, /cl/+/th/
F-score for M & S	0.63	**0.67**
F-score for I	0.33	**0.47**

5. Conclusion

- Singing has only a subset of the errors found in speech (consonant substitutions).
- We provided rules that predict singing mispronunciations for a given L1.
- Combining sub-phonetic American English models for approximating the missing phone models of L1 is useful.
- Our knowledge-based approach for singing pronunciation evaluation is promising.

* This work has been published in the proceedings of *ISMIR 2017,* "**Towards Automatic Mispronunciation Detection in Singing**", by Chitralekha Gupta, David Grunberg, Preeti Rao, and Ye Wang

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