Integrating Articulatory Features into Acoustic-Phonemic Model for Mispronunciation Detection and Diagnosis in L2 English Speech

***Shaoguang Mao*[1], *Zhiyong Wu*[1,2], *Xu Li*[2], *Runnan Li*[1], *Xixin Wu*[1], *Helen Meng*[1,2]**

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University

[2] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong

1. Introduction

Objective
- Mispronunciation detection and diagnosis (MDD) of L2 learner's speech

Challenge
- Pronunciations from L2-learners are nonstandard and unordered
- Traditional acoustic features cannot provide enough information about L2 learner's pronunciations

Motivation
- Pronunciation is decided by production position and mechanism, i.e. articulatory features
- Employ articulatory features to boost the MDD system
- Articulatory features can help L2 learners correct their mispronunciation

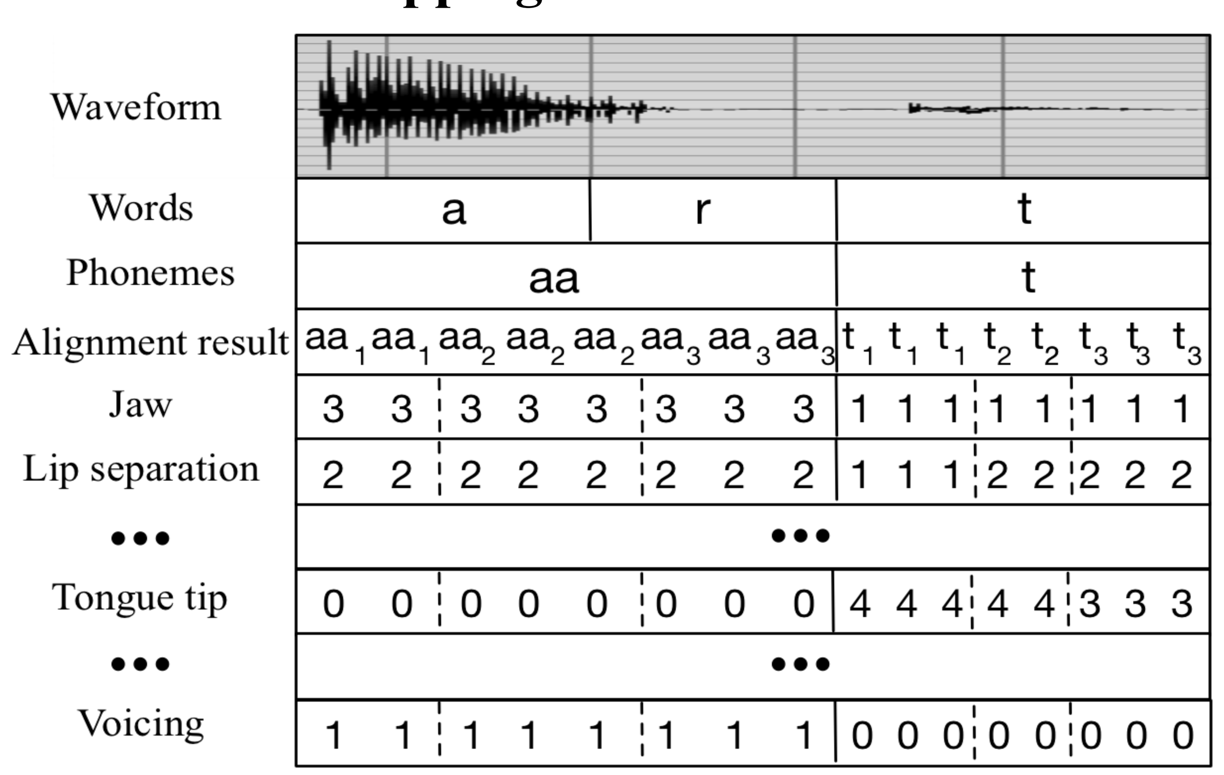
Contribution
- Introduce the articulatory features to MDD in deep learning framework
- Investigate several architectures for better exploiting articulatory features

2. Articulatory Features

Articulatory Feature Space [1]

stream	*classes*	*cardinality*
jaw	0: Nearly Closed, 1: Neutral, 2: Slightly Lowered, 3: Lowered	4
lip separation	0: Closed, 1: Slightly Apart, 2: Apart, 3: Wide Apart	4
lip rounding	0: Rounded, 1: Slightly Rounded, 2: Neutral, 3: Spread	4
tongue frontness	0: Back, 1: Slightly Back, 2: Neutral, 3: Slightly Front, 4: Front	5
tongue height	0: Low, 1: Mid, 2: Mid-High, 3: High	4
tongue tip	0: Low, 1: Neutral, 2: Dental, 3: Nearly Alveolar, 4: Alveolar	5
velum	0: Closed, 1: Open	2
voicing	0: Unvoiced, 1: Voiced	2

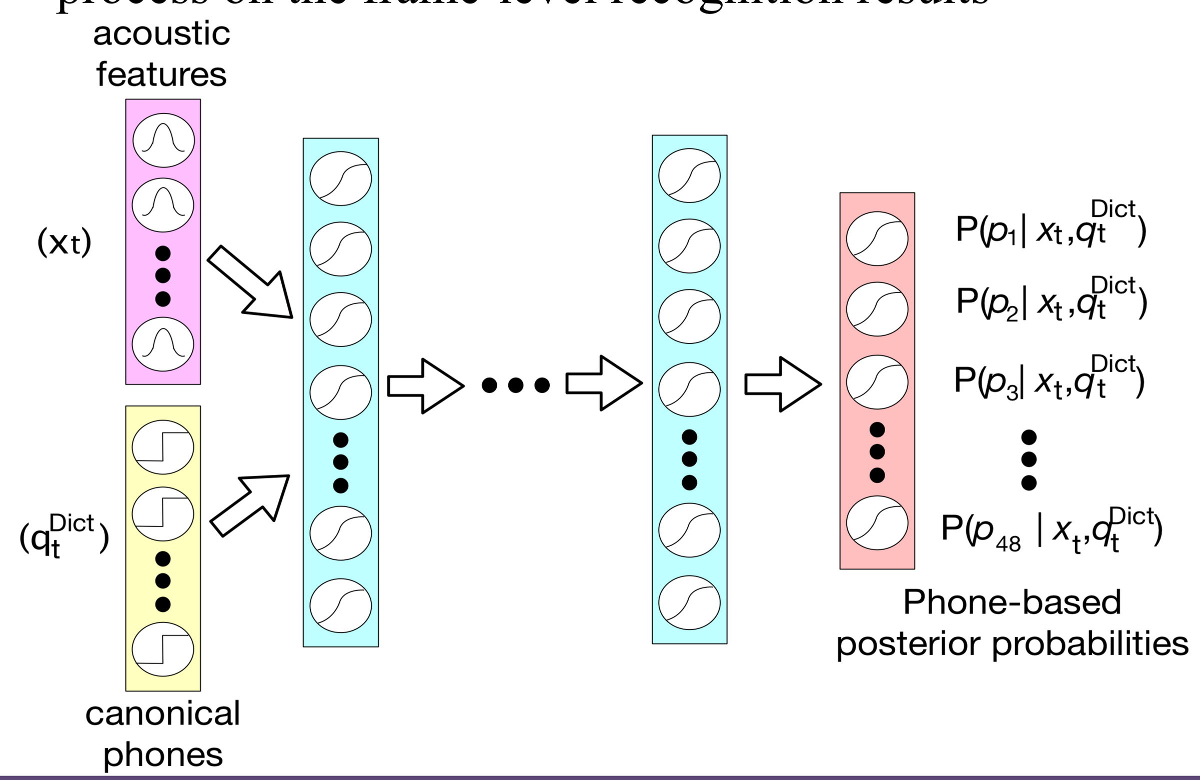
Phoneme Mapping

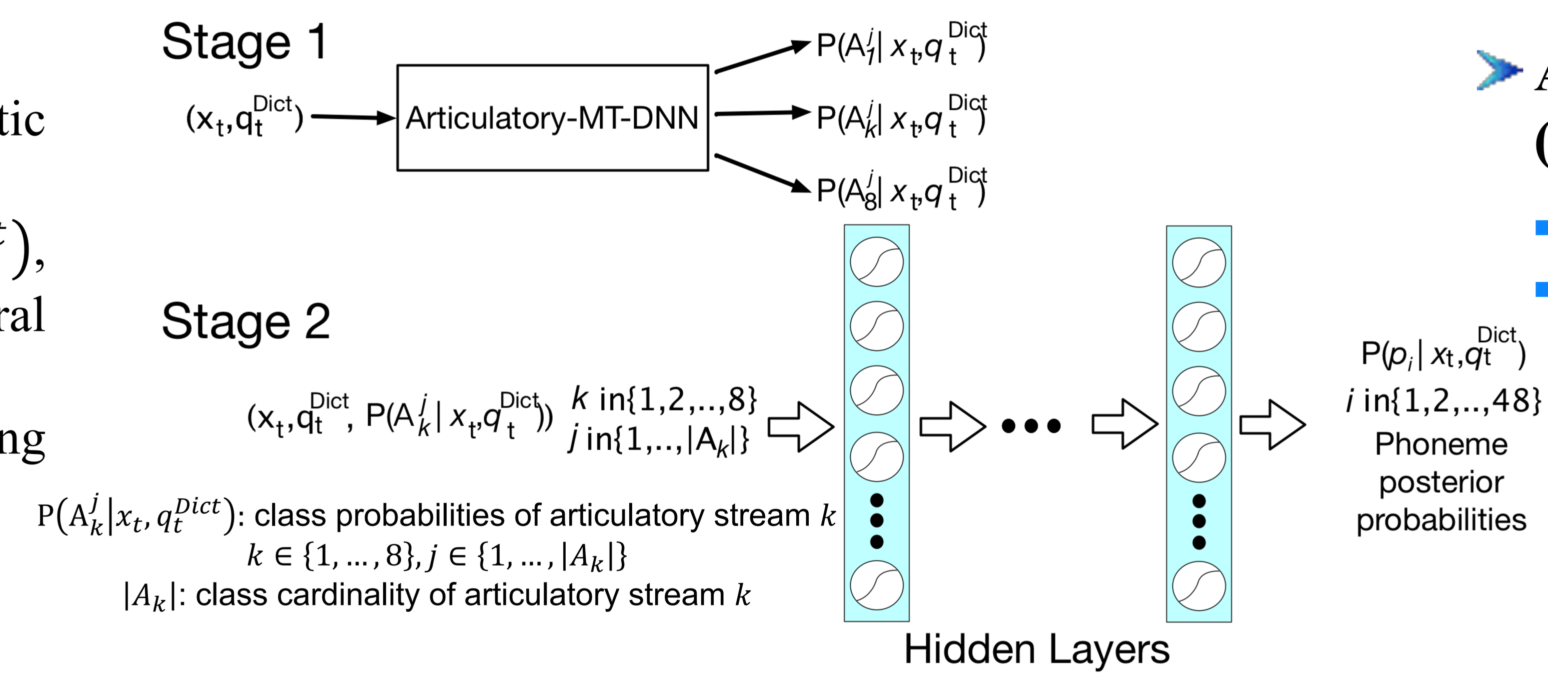


IPA	ɑː	t	
Phoneme	*aa*	*t*	
jaw	3	1	1
lip separation	2	1	2
lip rounding	1	2	2
tongue frontness	1	4	4
tongue height	0	3	2
tongue tip	0	4	3
velum	0	0	0
voicing	1	0	0

3. Acoustic-Phonemic Model with Articulatory Features

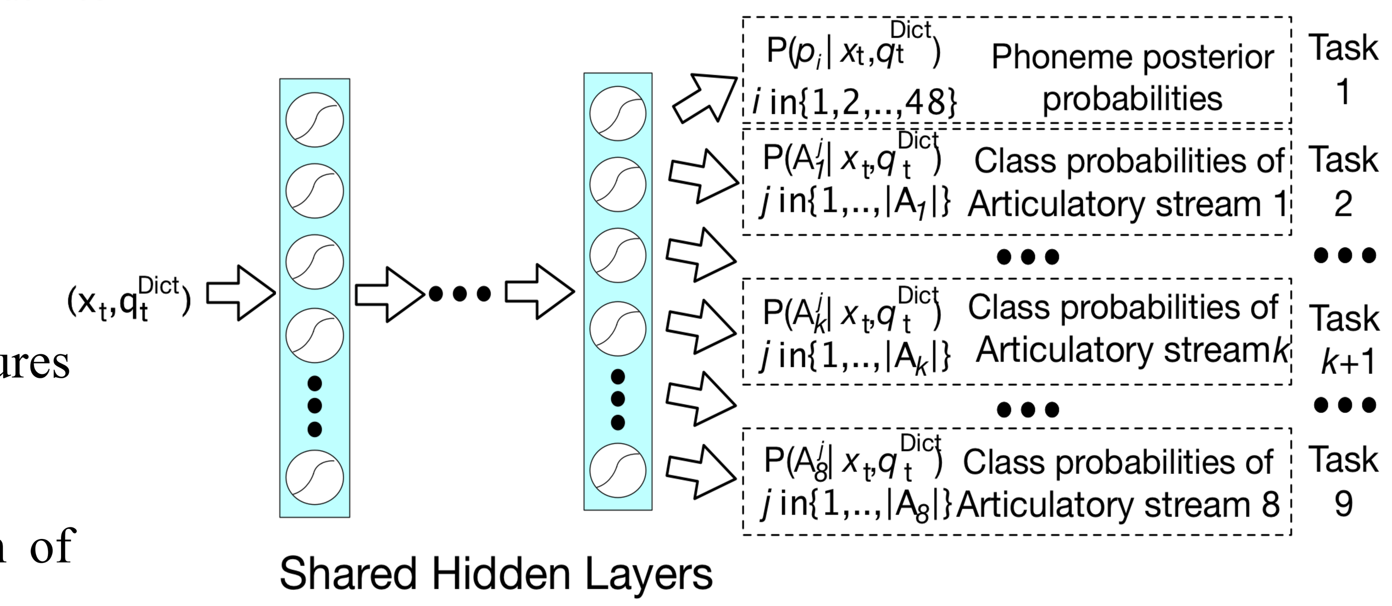
Acoustic-Phonemic Model (APM)
- Concatenate acoustic features (x_t, MFCC) and phonetic features (q_t^{Dict}, canonical phoneme) as input
- Derive phoneme posterior probabilities $P(p_i|x_t, q_t^{Dict})$, $i \in [1, \dots, 48]$ after several hidden layers in deep neural network (DNN)
- Generate recognized phone sequence by a smoothing process on the frame-level recognition results





Articulatory-Acoustic-Phonemic Model (AAPM)
- **Stage 1:**
  - Train an articulatory multi-task DNN to predict articulatory features
- **Stage 2**:
  - Embed the predicted articulatory features into input features
  - Derive phoneme posterior probabilities with the concatenation of x_t, q_t^{Dict} and the predicted articulatory features

Articulatory Multi-Task Acoustic-Phonemic Model (A-MT-APM)
- Incorporate the multi-task learning technique into APM
- Train the phoneme recognizer and articulatory stream classifiers together



4. Experiments & Conclusions

Corpus: CU-CHLOE (**C**hinese **U**niversity **Ch**inese **L**earners **o**f **E**nglish)

Experimental Setup
- Comparing models :
  1) State-based APM; 2) Phoneme-based APM; 3) AAPM;
  4) Bottleneck-AAPM; 5)A-MT-APM
- Acoustic features (x_t) : 11 frames (5 before, 1 current and 5 after) of MFCC
- Phonemic features (q_t^{Dict}) : 7 canonical phones (3 before, 1 current and 3after)

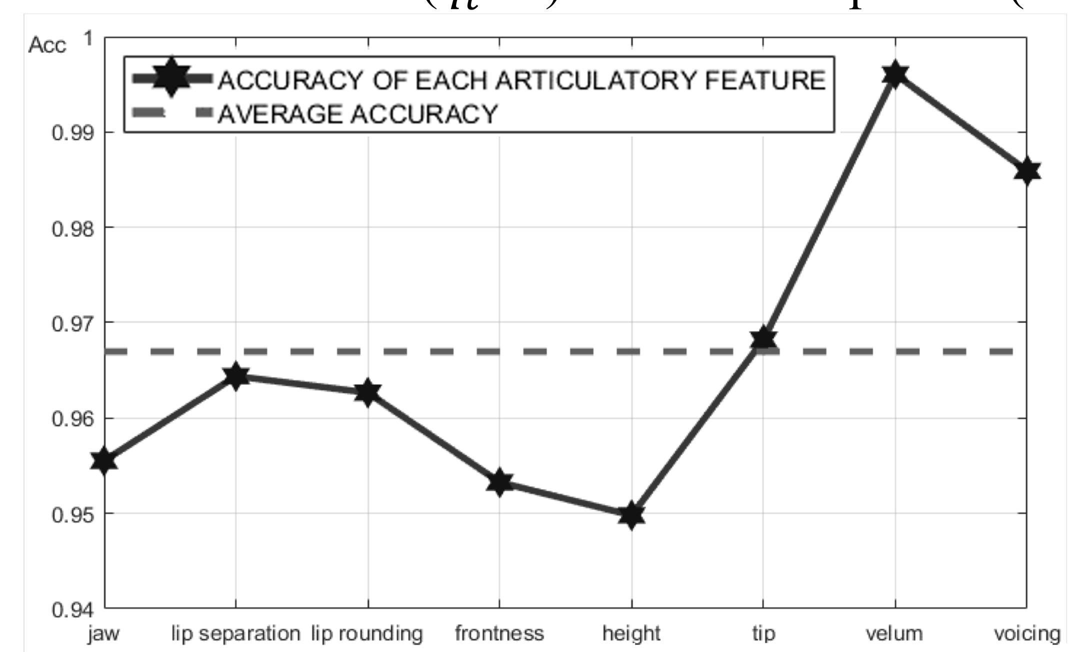


Fig. The accuracy of articulatory feature prediction

Evaluation Metrics and Experimental Results

$$Precision = \frac{TR}{TR+FR}, \quad Recall = \frac{TR}{TR+FA}$$

$$F-measure = 2\times\frac{Precision\times Recall}{Precision + Recall}$$

$$Detection\ Accuracy = \frac{TA+TR}{TA+FR+FA+TR}$$

$$Diagnostic\ Accuracy = \frac{CD}{CD+DE}$$

Table.1 The definitions in hierarchical evaluation

		Recognition Result	
		C.P.	M.P.
Manually Transcribed Phonetic Unit	C.P.	**TA**	**FR**
	M.P.	**FA**	**TR (CD/DE)**

C.P.: Correct pronunciation; M.P.: Mispronunciation.

Table.2 Experimental results of MDD with different metrics

Method	*Performance of Mispronunciation Detection and Diagnosis*				
	Precision	*Recall*	*F1-measure*	*Detection Accuracy*	*Diagnosis Accuracy*
State-based APM	63.4%	**83.7%**	72.1%	89.0%	68.4%
Phoneme-based APM	71.8%	80.2%	75.7%	92.1%	77.3%
AAPM	72.9%	80.9%	76.7%	92.5%	77.4%
Bottleneck-AAPM	79.8%	77.9%	78.8%	93.7%	80.4%
A-MT-APM	**86.5%**	76.7%	**81.3%**	**94.6%**	**84.3%**

Conclusions:
- Reliable articulatory features can be predicted with acoustic and phonemic inputs
- APM with articulatory features greatly improves the MDD performance, especially when the phoneme recognizer are jointly trained with articulatory stream classifiers

5. Acknowledgment

- This project is partially supported by a grant from the HKSAR RGC General Research Fund (project no. 14207315), and a seed grant from the MSRA Collaborative Research Project.
- The research was conducted while the first author was an intern at CUHK.

[1] J. Tepperman and S. Narayanan. "Using articulatory representations to detect segmental errors in nonnative pronunciation." *IEEE TASLP, 16(1): 8-22, 2008.*