APPLYING MULTITASK LEARNING TO ACOUSTIC-PHONEMIC MODELING FOR MISPRONUNCIATION DETECTION AND DIAGNOSIS IN L2 ENGLISH SPEECH

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

[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
- Unbalanced data distribution between correct and incorrect L2 speech
- Existing approaches insufficiently capture differences in between correct and incorrect phoneme pronunciations

Multi-Task Training
- Process correct and incorrect pronunciations in L2 speech separately
- Correct-pronunciation recognizer to focus on correct pronunciation
- Mispronunciation recognizer to focus on incorrect pronunciations
- Train two tasks together with multi-task learning

Contribution
- Propose multi-task Acoustic-Phonemic Model (MT-APM) and related feature representation (R-MT-APM)

2. Acoustic-Phonemic Model

Input Features
- Concatenate acoustic features (x_t, i.e. MFCC) and phonetic features (q_t^{Dict}, i.e. current canonical phone with 3 phones to the left and right respectively)

Structure (Fig. 1)
- Derive phone-state posterior probabilities $\mathrm{P}(s_i|x_t,q_t^{Dict}), i \in [1,\dots,144]$ after several hidden layers;
- Generate recognized phone sequence with Viterbi decoding

Problems
- Low recall (<70%) of mispronounced phones

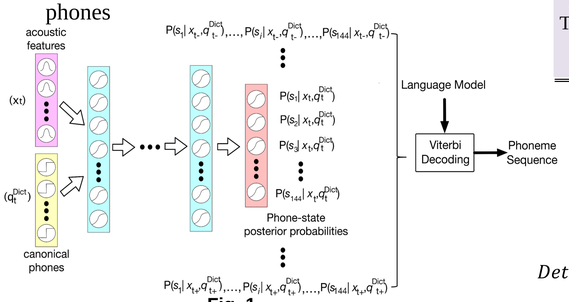

Fig. 1

3. Multi-Task Learning for APM

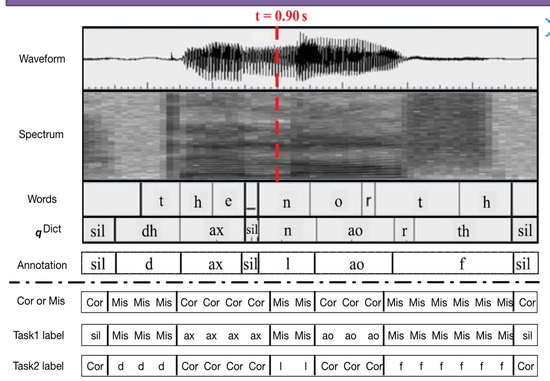

Fig. 2

Data Labeling (Fig. 2)
- Introduce two new states (***mis*** and ***cor***) for the two tasks in (R-)MT-APM
  - For a frame, compare its annotation with canonical phone;
  - If same (correct pronunciation), its label for Task 1 is the canonical phone state $cs_i, i \in [1 \dots 144]$, while its label for Task 2 is ***cor*** ;
  - If different (mispronunciation), its label for Task 1 is ***mis***, while its label for Task 2 is the annotation phone state $ms_i, i \in [1 \dots 144]$.

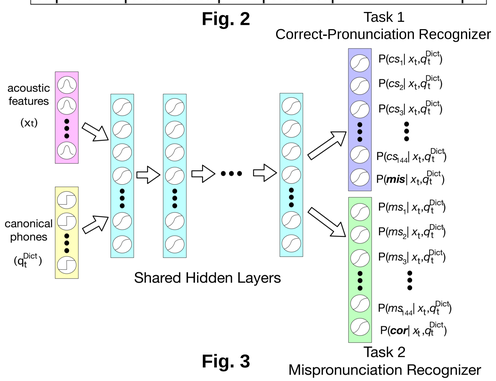

Fig. 3

Multi-Task APM (MT-APM)
- **Multi-task Structure (Fig. 3)**
  - Task 1 : Correct-pronunciation Recognizer
  - Task 2 : Mispronunciation Recognizer
  - Train two tasks together with multi-task learning
- **Joint Decoding for MT-APM (Fig. 4)**
  - Compare $\mathrm{P}(mis|x_t,q_t^{Dict})$ from Task 1 and $\mathrm{P}(cor|x_t,q_t^{Dict})$ from Task 2.
  - Use $\mathrm{P}(ms_i|x_t,q_t^{Dict}), i \in [1 \dots 144]$ from task 2 as the output for Viterbi decoding if $\mathrm{P}(mis|x_t,q_t^{Dict}) > \mathrm{P}(cor|x_t,q_t^{Dict})$.
  - Else, use $\mathrm{P}(cs_i|x_t,q_t^{Dict}), i \in [1 \dots 144]$ for decoding.

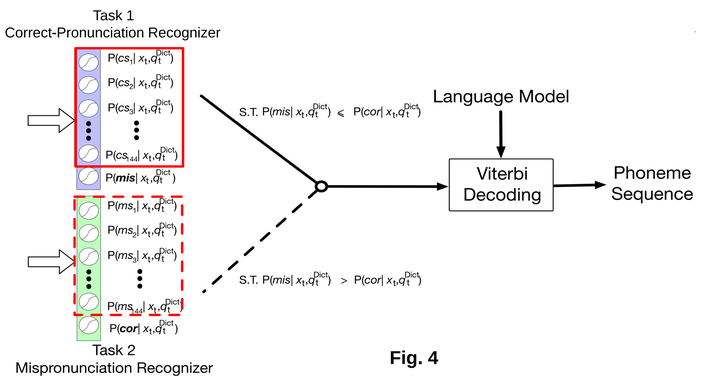

Fig. 4

Feature Representation for MT-APM (R-MT-APM)
- **Two Stage Structure (Fig. 5)**
- **Stage 1**
  - Train correct-mispronunciation DNN (CM-DNN) to judge whether current frame is ***cor*** or ***mis***;
- **Stage 2**
  - Train the dense layer and shared hidden layers with the fixed pre-trained CM-DNN;
  - Derive $\mathrm{P}(C|x_t,q_t^{Dict})$ and $\mathrm{P}(M|x_t,q_t^{Dict})$ for input features;
  - Compute a dense output vector;
  - Compute the represented new features by adding input features and the dense output vector.

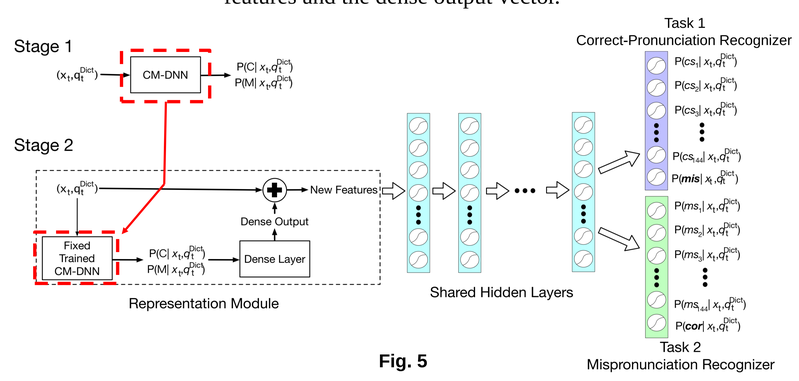

Fig. 5

4. Experiments

Corpus: CU-CHLOE (**C**hinese **U**niversity **Ch**inese **L**earners **o**f **E**nglish)
- L2 English speech uttered by 100 Cantonese speakers (CHLOE-C)
- 30% speaker audios are labeled by skilled linguists with actual pronunciations

Evaluation Metrics and Experimental Results

		Recognition Result	
		Correct Pronunciation	Mispronunciation
Manually Transcribed Phonetic Unit	Correct Pronunciation	**TA**	**FR**
	Mispronunciation	**FA**	**TR (CD/DE)**

- The Recognition metrics:

$$Correct = \frac{N-S-D}{N}, \quad Accuracy = \frac{N-S-D-I}{N}$$

- The MDD metrics :

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

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

$$Detection\ Accuracy = \frac{TA+TR}{TA+FR+FA+TR}, \quad Diagnostic\ Accuracy = \frac{CD}{CD+DE}$$

Experimental Setup
- Comparing models :

  1) Baseline APM; 2) MT-APM; 3) R-MT-APM; 4) A-MT-APM.
- Acoustic features (x_t) : 11 frames (5 before, 1 current and 5 after) of MFCC, using 25-ms Hamming window and 10-ms frame shift
- Phonemic features (q_t^{Dict}) : 7 canonical phones (3 before, 1 current and 3after)

Dataset	Method	*Performance of Recognition*		*Performance of Mispronunciation Detection and Diagnosis*				
		Correct	*Accuracy*	*Precision*	*Recall*	*F-measure*	*Detection Accuracy*	*Diagnostic Accuracy*
Small Scale (5h)	APM	79.60%	72.20%	52.02%	84.67%	64.44%	84.24%	57.07%
	MT-APM	84.40%	76.80%	59.31%	**89.33%**	71.29%	87.86%	**75.69%**
	R-MT-APM	**86.10%**	**77.00%**	**63.47%**	88.78%	**74.02%**	**89.44%**	74.07%
	A-MT-APM	74.80%	61.80%	52.02%	84.67%	64.44%	84.24%	57.07%
Medium Scale (7.5h)	APM	80.80%	78.60%	53.22%	83.56%	65.02%	84.92%	73.47%
	MT-APM	85.50%	81.70%	61.29%	86.14%	71.62%	88.54%	75.83%
	R-MT-APM	**87.50%**	**83.10%**	**65.84%**	89.92%	**76.02%**	**90.44%**	**77.71%**
	A-MT-APM	83.70%	78.70%	62.26%	**90.35%**	73.72%	89.10%	73.77%
Large Scale (9.5h)	APM	81.40%	76.30%	63.35%	83.74%	72.13%	89.03%	68.36%
	MT-APM	86.40%	80.50%	62.78%	89.05%	73.64%	89.26%	**79.63%**
	R-MT-APM	**88.20%**	**83.30%**	67.65%	**89.52%**	**77.07%**	**90.99%**	78.24%
	A-MT-APM	86.80%	81.30%	**67.75%**	85.60%	75.63%	90.70%	75.72%

5. Conclusion
- Propose MT-APM and R-MT-APM
- Better capture differences in between correct and incorrect phoneme pronunciations
- Resolve the low recall problem in MDD

6. 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.