UNSUPERVISED DISCOVERY OF AN EXTENDED PHONEME SET IN L2 ENGLISH SPEECH FOR MISPRONUNCIATION DETECTION AND DIAGNOSIS

***Shaoguang Mao*[1,2], *Xu Li*[1], *Kun Li*[1] , *Zhiyong Wu*[1,2,*], *Xunying Liu*[1], *Helen Meng*[1,2]**

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

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

INTRODUCTION

Problem Statement

- Detect mispronunciations in second language (L2) speech and provide diagnostic feedback
- Existing approaches mainly target categorical phonetic errors based on native (L1) phoneme set and cannot handle non-categorical errors appropriately
- Goal: Discover Extended Phoneme Set in L2 speech (L2-EPS), covering both categorical and non-categorical phoneme units

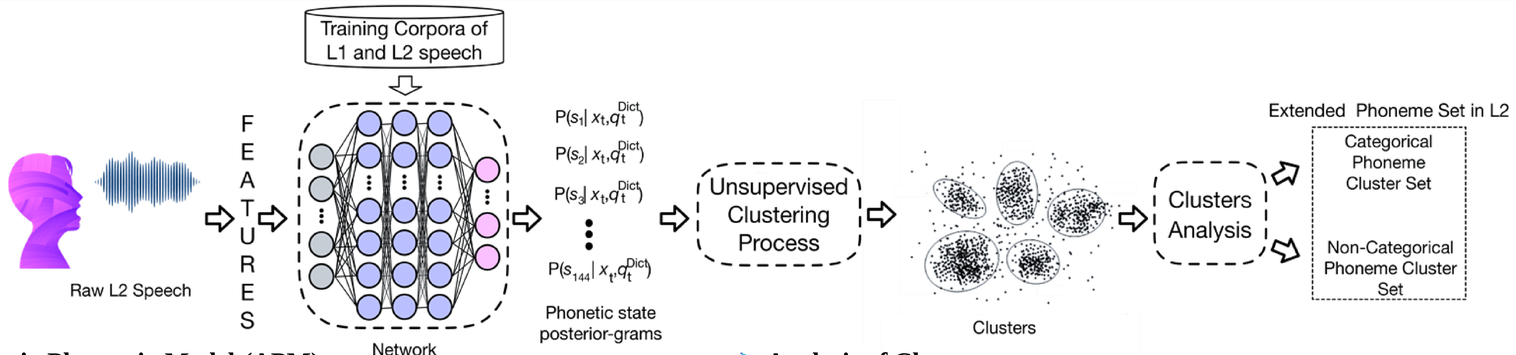
Proposed Approach

- Phonetic Posterior-Grams (PPGs) to represent L2 English acoustic-phonetic space
- Unsupervised clustering of L2 speech frames with PPG features to uncover potential phoneme patterns
- Analyze clusters and label as categorical/non-categorical phonemes

Word	n o r th		
Canonical Text	n ao r th		
Real Pronunciation	n_l ao r th		
Traditional Annotation	n ao r th		
Recognition Result 1	l ao r th	✓	✗
Recognition Result 2	n ao r th	✗	✗

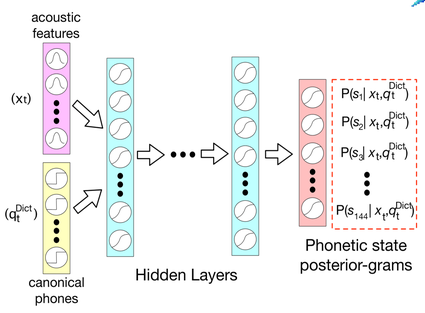
Detection | Diagnosis

APPROACH



Acoustic-Phonemic Model (APM) Generating Phonetic Posterior-Grams (PPGs)

- Acoustic features (x_t) : Mel-frequency cepstral coefficients
- Phonemic features (q_t^{Dict}) : 7 canonical phones (3 before, 1 current and 3 after)
- Phonetic state posteriorgrams: vectors of posterior probabilities of each phonetic unit, used for clustering



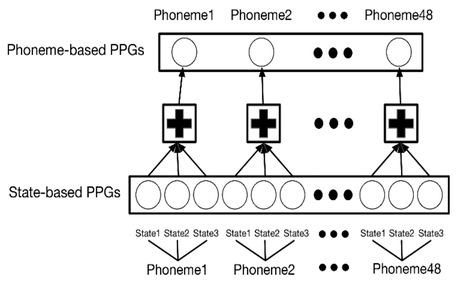
Unsupervised Clustering Process

- Transform state-based PPGs (v_s) into phoneme-based PPGs (v_p): $Av_s = v_p$

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & \cdots & 0 \\ \vdots & & \cdots & & \ddots & & \cdots & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

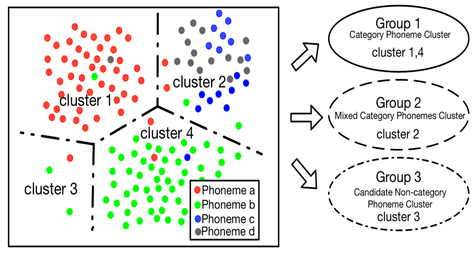
, $A \in R^{48\times144}, v_s \in R^{144}, v_p \in R^{48}$

- Perform n-best filtering on the phoneme-based PPG
- K-means clustering with random initialization



Analysis of Clusters

- Group clusters based on mapping with L1 phonemes
- An L1 phoneme maps to a cluster if proportion of speech frames labeled with the Phoneme $P(N_p)$ and also classified into Cluster $i(N_{pi})$ exceeds threshold δ (set at 90%), i.e. $\frac{N_{pi}}{N_p} \geq \delta$
- Analyze clusters by centroids
  - ❖ Focus on each cluster centroid as the representative average PPG
    - ✓ Categorical phoneme clusters: cluster centroids have only one peak which appears at the related L1 phoneme bit
    - ✓ Non-categorical phoneme clusters: cluster centroids may have multiple peaks appearing at different L1 phoneme bits
  - ❖ For non-categorical phoneme clusters, label them with m phonemes, where the top m peaks appear



L1 data in different clusters

	Description	Requirement
Group 1	Categorical Phoneme Clusters	Only one categorical phoneme maps to this cluster
Group 2	Mixed Categorical Phonemes Clusters	More than one categorical phoneme maps to this cluster
Group 3	Candidate Non-categorical Phoneme Clusters	Clusters not in Group 1 or Group 2

EXPERIMENTS

Corpus

- L1 corpus : TIMIT (about 4 hours)
- L2 corpus : CU-CHLOE (Chinese University-Chinese Learners of English)
  - ❖ L2 English speech uttered by 100 Cantonese speakers (CHLOE-C) (about 12 hours)
  - ❖ 30% speaker audios are labeled by skilled linguists with categorical phonemes

Setup

- k value in *k-means* : from 70 to 120 with step-length being 10
- n-best filtering : n = 3
- Network configuration:

	No. of hidden layers	No. of units per layer	Activation function
APM&DNN	5	2048	tanh
LSTM	2	512	tanh

Clustering Setups and Evaluation

- Frame-level features for clustering :
  - ❖ MFCC; ❖ State-level PPGs (derived from DNN, LSTM and APM);
  - ❖ Phoneme-level PPGs (derived from DNN, LSTM and APM).
- Davies Bouldin Index (DBI)

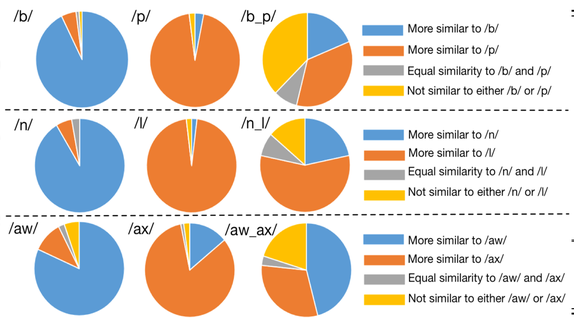
$$DBI \equiv \frac{1}{N}\sum_{i=1}^{N} max_{j\neq i}\frac{S_i + S_j}{d_{i,j}}$$

$$S_i = \frac{1}{|C_i|}\left(\sum_{X\in C_i}\|X - Z_i\|\right), d_{ij} = \|Z_i - Z_j\|$$

Features	MFCC	PPGs from DNN		PPGs from LSTM		PPGs from APM	
		State-based	Phoneme-based	State-based	Phoneme-based	State-based	Phoneme-based
k = 70	2.17	1.87	1.62	1.77	1.61	1.53	1.34
k = 80	2.19	1.91	1.57	1.76	1.59	1.57	1.33
k = 90	2.17	1.94	1.60	1.77	1.61	1.49	1.28
k = 100	2.16	1.86	1.58	1.84	1.55	1.35	**1.26**
k = 110	2.19	1.92	1.56	1.74	1.51	1.49	1.29
k = 120	2.18	1.93	1.55	1.67	1.50	1.60	1.43

PERCEPTUAL TESTS

- For each non-categorical phoneme comparison, 30 audio files randomly played
  - ❖ Non-categorical cluster ➔ 10 audio files
  - ❖ Related categorical Cluster 1 and 2 ➔ 10 audio files for each
  - ✓ 1) More similar to P_1
  - ✓ 2) More similar to P_2
  - ✓ 3) Equal similarity to P_1 and P_2
  - ✓ 4) Not similar to either P_1 or P_2



Categorical Phonemes	sil	aa	ae	ah	ao	aw
	ax	ay	b	ch	cl	d
	dh	dx	eh	el	en	epi
	er	ey	f	g	hh	ih
	ix	iy	jh	k	l	m
	n	ng	ow	oy	p	r
	s	sh	t	th	uh	uw
	v	vcl	w	y	z	zh
Non-categorical Phoneme	aa_ao	aa_ax	ae_ay	ae_ay	ax_er	b_p
	eh_ey	ey_ih	f_v	m_n	n_l	t_d

CONCLUSION

- Propose a framework to discover L2 Extended Phoneme Set
- Improve coverage of pronunciation patterns in L2 speech
- To be incorporated with existing MDD approaches for better performance

ACKNOWLEDGEMENT

- This project is partially supported by a grant from the HKSAR RGC General Research Fund (project no. 14207315) and a seed grant of the Microsoft Collaborative Research Project.