UNSUPERVISED DISCOVERY OF NON-NATIVE PHONETIC PATTERNS IN L2 ENGLISH SPEECH FOR MISPRONUNCIATION DETECTION AND DIAGNOSIS

***Xu Li*[1], *Shaoguang Mao*[2], *Xixin Wu*[1], *Kun Li*[3], *Xunying Liu*[1], *Helen Meng*[1,2,*]**

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

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

[3]SpeechX Limited, Shenzhen

INTRODUCTION

Problem Statement

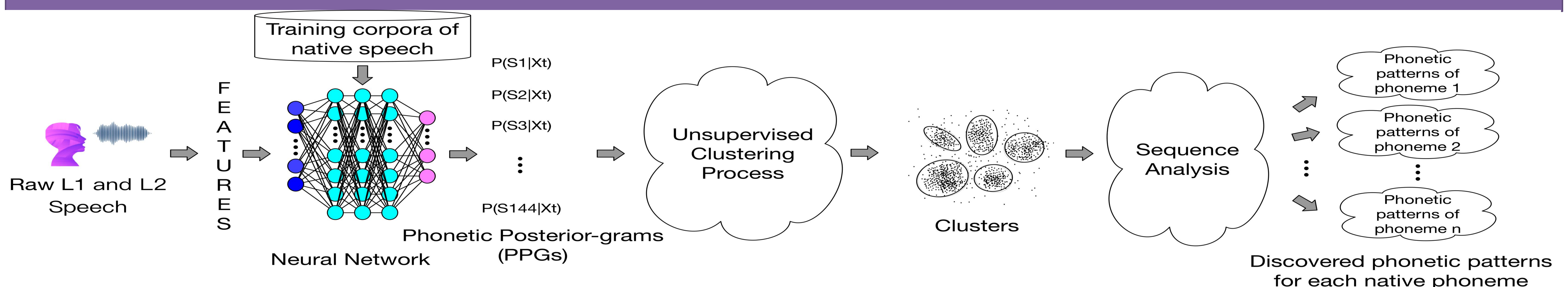
- Detect mispronunciations in second language (L2) speech and provide diagnostic feedback
- Existing approaches mainly target categorical phonetic errors based on native (L1) phoneme set but cannot handle non-native phonetic patterns occurred in L2 speech
- Goal: Discover non-native phonetic patterns of each native phoneme to better cover the pronunciation patterns in L2 speech

Proposed Approach

- Phonetic Posterior-Grams (PPGs) to represent L2 English acoustic-phonetic space
- Unsupervised clustering of L2 speech frames based on PPGs and use cluster ID sequence to represent segment level information
- Apply Cluster Sequence Analysis (CSA) to discover each phoneme's potential non-native phonetic patterns

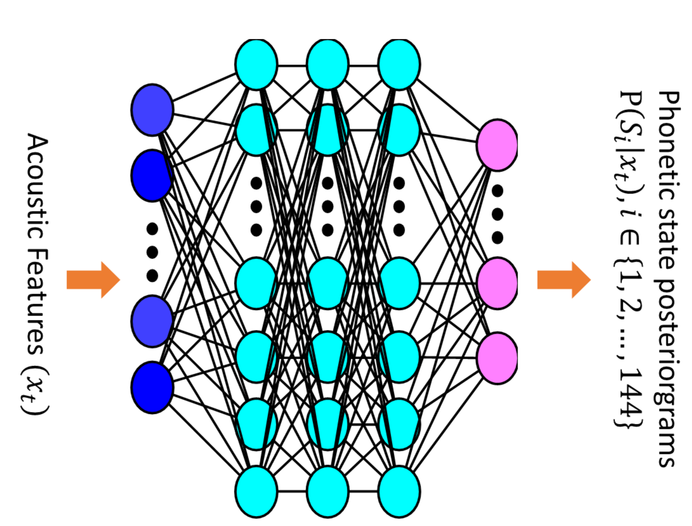
		Detection	Diagnosis
Word	hate		
Canonical Text	hh ey t		
Real Pronunciation	hh ey t_s		
Traditional Annotation	hh ey t		
Recognition Result 1	hh ey s	Correctly Detected	Wrong
Recognition Result 2	hh ey t	Missed	Wrong

APPROACH FRAMEWORK



Neural Network Model Generating Phonetic Posterior-Grams (PPGs)

- Acoustic features (x_t) : Mel-frequency cepstral coefficients
- Phonetic state posteriorgrams: vectors of posterior probabilities of each phonetic unit, used for clustering
- Neural Network Type: Simple deep neural network with fully connected layers

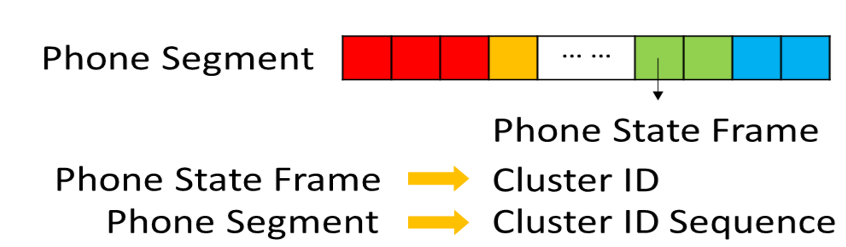


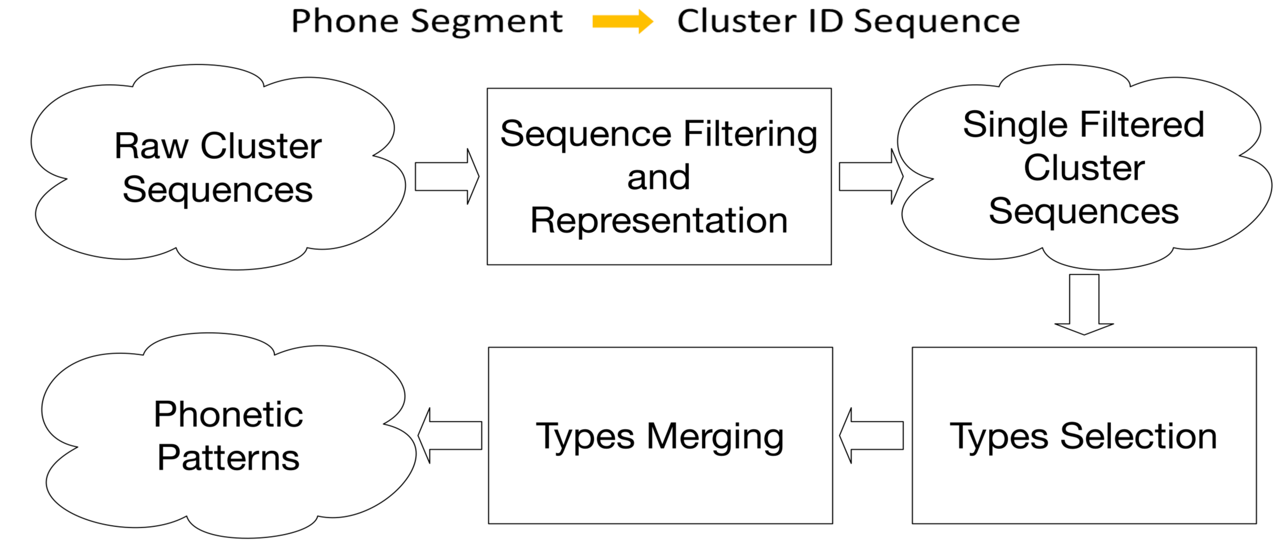
Cluster Sequence Analysis (CSA)

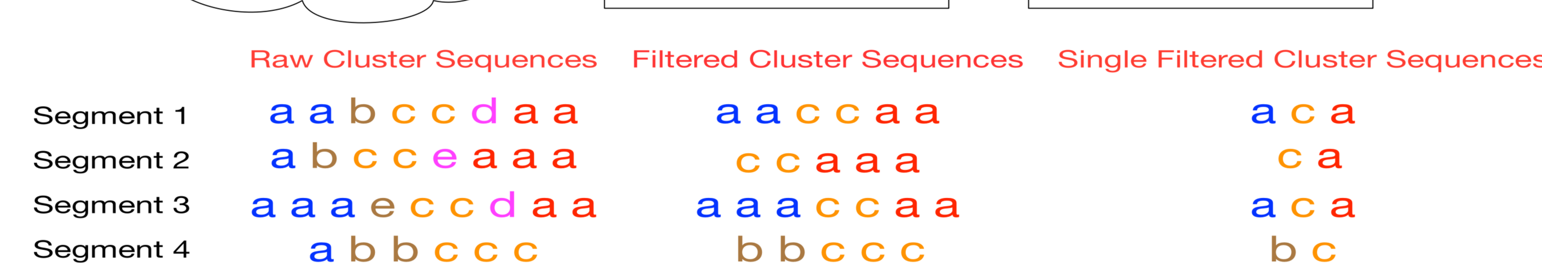
- Preprocessing
  - Group phone segments by native phonemes according to forced alignment
- Algorithm
  - Sequence Filtering and Representation
    - Remove cluster IDs that are very transient (the length of continuous frames belonging to a cluster ID is below a threshold, e.g. # frames = 2): Raw Cluster Sequence ➡ Filtered Cluster Sequence (FCS)
    - Preserve a single cluster ID among a continuous pieces of a same cluster ID: FCS ➡ Single Filtered Cluster Sequence (SFCS)
  - Type Selection
    - Preserve the SFCS with the frequency above a threshold, e.g. F = 0.05
  - Type Merging
    - Merge a type into another when it is a sub-sequence of the second one (the phonetic deviations between these two types are relatively small) : SFCS ➡ Merged SFCS (MSFCS)
- Apply CSA in exploring phonetic patterns
  - Apply CSA to L1 speech ➡ Canonical phonetic patterns
  - Apply CSA to L2 speech ➡ Non-native phonetic patterns

Unsupervised Clustering Process

- Capture the variation of state-level phonetic patterns in L2 speech
- Perform n-best filtering on state-level PPGs
- K-means clustering with random initialization
- A cluster ID to represent each phone state frame, and further a phone segment representation consists of a cluster ID sequence







"SFCS" Types	Frequency	"MSFCS" Types	Frequency
aca	2/4	aca	3/4
ca	1/4	bc	1/4
bc	1/4		

EXPERIMENTS

Corpus

- L1 corpus : TIMIT (about 5 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 111 to 234 with step-length being 3
- *n*-best filtering : $n = 3$
- Network configuration: 4 hidden layers with 1024 units per layer and tanh as activation function

	DNN Model Training		Unsupervised Clustering
	Train Set	Dev Set	Test Set
L1 corpus	3.17 hours	0.5 hours	1.33 hours
L2 corpus			3.6 hours

Clustering Setups and Evaluation

- Frame-level features for clustering :
  - MFCC;
  - State-level PPGs derived from DNNs;
- 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	MFCCs	PPGs from DNN
K=111	2.21	1.88
K=123	2.20	1.84
K=135	2.19	1.94
K=147	2.17	1.79
K=159	2.19	1.76
K=171	2.18	1.75
K=**174**	2.19	**1.68**
K=183	2.21	1.85
K=195	2.21	1.86
K=207	2.22	1.93
K=219	2.22	1.94
K=231	2.20	1.90

PERCEPTUAL TESTS

- For each pair of phonetic patterns of a given phoneme
  - 5 audio files are randomly selected from each pattern and totally10 audio files are displayed
  - Subjects are asked -- "Are the phonetic patterns in the two audios the same or not?"

✓ 1) Yes ✓ 2) No ✓ 3) Not Sure

Deviations between canonical and non-native phonetic patterns of example phonemes

	ae	aw	ax	eh	ey	f	ix	iy	jh	t
Same	26.5%	35.8%	25.3%	29.0%	40.7%	44.4%	32.7%	42.1%	34.6%	34.6%
Different	66.7%	58.0%	69.8%	62.3%	53.1%	49.4%	61.1%	49.9%	61.7%	57.4%
Not Sure	6.8%	6.2%	4.9%	8.6%	6.2%	6.2%	6.2%	8.0%	3.7%	8.0%

Statistical results of canonical and non-native phonetic patterns among all native phonemes

	Yes	No	Not Sure
Mean	37.6%	55.9%	6.5%
std	0.109	0.104	0.015

CONCLUSION

- Proposed a framework to discover non-native patterns given a native phoneme
- Seek to improve coverage of pronunciation patterns in L2 speech
- To be incorporated into mispronunciation detection and diagnosis in L2 learner's speech

ACKNOWLEDGEMENT

- This project is partially supported by a grant from the HKSAR RGC General Research Fund (project no. 14207315).