# Computer-Assisted Pronunciation Training

Dong  Wenwei
2019.1.17

# Outline

- Introduction
- Mispronunciation patterns of non-native speakers
- Research approaches
- Challenges and research opportunities

# Introduction

- With increasing globalization, there has also been a significant increase in the demand for foreign language learning

分时共用的教师资源

# Introduction

- CAPT特点
  - 方便学生进行大量发音练习和测试
  - 个性化学习进度
  - 方便教师掌握学生发音情况
  - 提升发音评测的客观性

# Introduction

- Applications of CAPT can be divided into two areas:
  - Pronunciation assessment
  - Pronunciation learning/teaching
    - Segmental (phonetic)
    - Subsegmental (e.g., place of articulation, manner of speech)
    - Suprasegmental (prosodic)

# Mispronunciation patterns of non-native speakers

- Pronunciation errors are usually characterized at the phonetic(segmental) or prosodic (suprasegmental) level
  - Phonetic Errors
    - Substitutions
    - Insertions
    - Deletions

# Mispronunciation patterns of non-native speakers

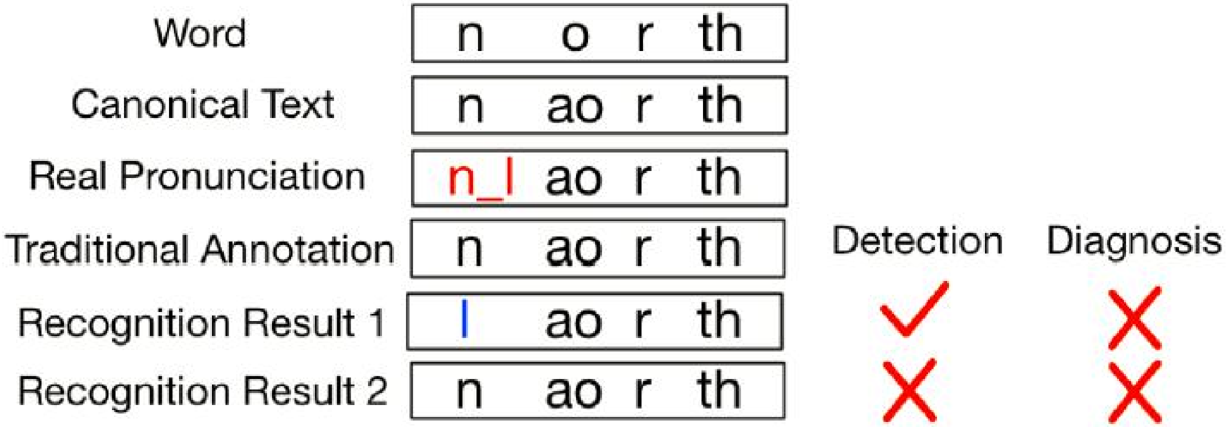
- Different phontactic constraints across languages might result in deletion and insertion errors
    - Only certain consonants are allowed at syllable final positions
        - “face” might be pronounced as “fay”
    - Consonant clusters are not allowed in Vietnamese either
        - Vowels might be inserted in between consonants when Vietnamese speakers learn English

# Mispronunciation patterns of non-native speakers

- Phonetic substitutions occur because of approximating L2 phonemes with L1 phonemes
    - In Mandarin and Spanish, there are no short vowels
        - Words like “eat” and “it” might sound similar
- Sometimes the non-native phone is neither in L1 or L2. It could be in between

# Mispronunciation patterns of non-native speakers

- Most existing approaches to modeling L2 speech can only target categorical phoneme error types based on the native phoneme set



|  |  |  |  |  | Detection | Diagnosis |
|---|---|---|---|---|---|---|
| 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 | ✗ | ✗ |

**Fig. 1.** An example for how non-categorical mispronunciations are wrongly treated in traditional MDD

# Mispronunciation patterns of non-native speakers

- In terms of intelligibility, prosody is as important as phonetic accuracy
  - Prosodic Errors
    - Stress
    - Rhythm
    - Intonation

# Mispronunciation patterns of non-native speakers

- Stress: the specific emphasis given to a particular syllable or word
  - Acoustic： greater loudness, higher pitch, and longer duration
  - The stress placed on syllables within words are called lexical stress or word stress
  - Stress placed on words within sentences are called sentence stress or prosodic stress

# Mispronunciation patterns of non-native speakers

- In Bengali（孟加拉语） is fixed (restricted to the initial syllable of a word)

- English has variable stress

# Mispronunciation patterns of non-native speakers

- Rhythm: the temporal pattern of how a language is spoken
  - English and German are stress-timed
    - Some syllables are long while others (unstressed syllables) are short
  - French and Spanish are syllable-based
    - Each syllable is spoken at a regular interval

# Mispronunciation patterns of non-native speakers

- Intonation: the variation in pitch
  - Intonation helps the listener parse the boundaries in speech
  - Intonation also helps convey the speaker’s attitude and emotions
- tonal languages such as Mandarin Chinese and Vietnamese
  - Variation in pitch can result in words with different meanings

# Research approaches

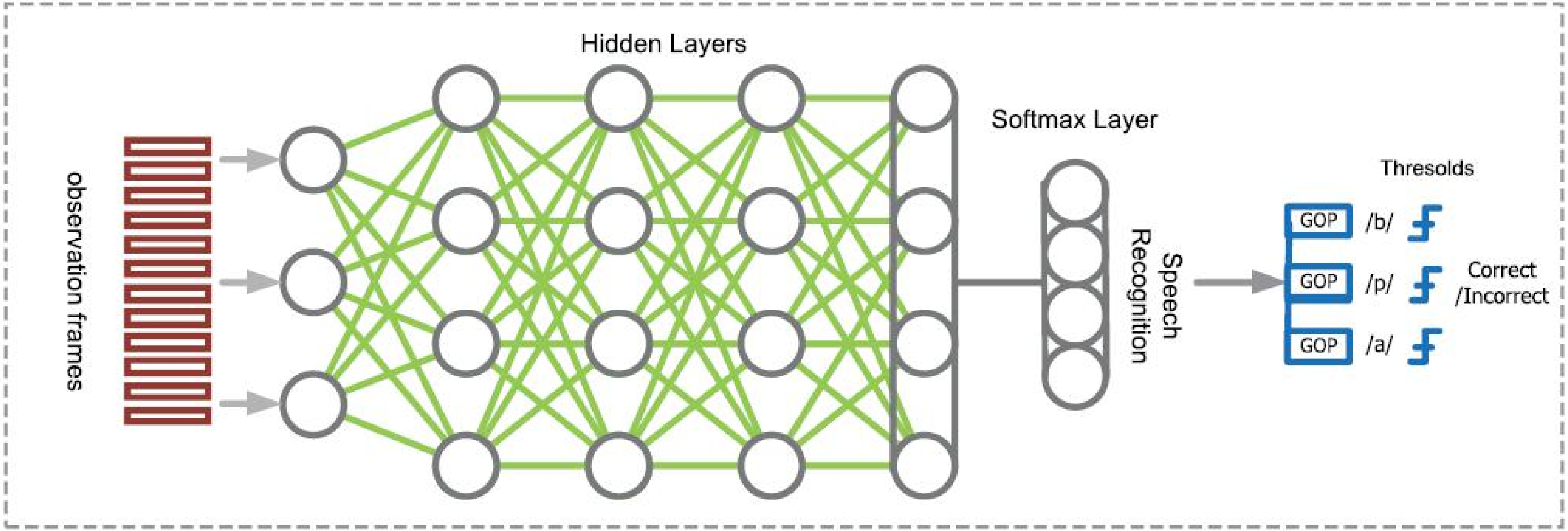
- Frameworks for Detecting Phonetic Errors：
  - ASR is often a natural component in a CAPT system
  - The ASR system can be trained with just native speech or with both non-native speech and native speech

# Research approaches

- Frameworks for Detecting Phonetic Errors：
  - Likelihood-Based Scoring（GOP）
  - Classifier-Based Scoring
  - Extended Recognition Network（ERN）
  - Unsupervised Error Discovery

# Research approaches

- Likelihood-Based Scoring（GOP）



# Research approaches

- Likelihood-Based Scoring（GOP）

$$\mathrm{GOP}(\mathrm{p}) = \log \frac{\mathrm{P}(\mathrm{O} \mid \mathrm{p})}{\max_{\mathrm{q}\in\mathrm{Q}} \mathrm{P}(\mathrm{O} \mid \mathrm{q})}$$

Q为所有音素的集合，p标准phone，O为声学特征
q后验概率最大的phone

$$\mathrm{GOP}(\mathcal{O}_n, q_n) > b \begin{cases} \text{yes,} & \text{correct pronunciation,} \\ \text{no,} & \text{mispronunciation.} \end{cases}$$

# Research approaches

- Classifier-Based Scoring
  - Truong et.al used acoustic phonetic features to train binary classifiers to distinguish confusion pairs
  - Acoustic phonetic，MFCC, GOP

# Research approaches

- Extended Recognition Networks(ERN)
  - 在解码网格中加入先验知识的约束(Kenworthy, 1987; A. M. Harrison, 2008;Gao, 2015)

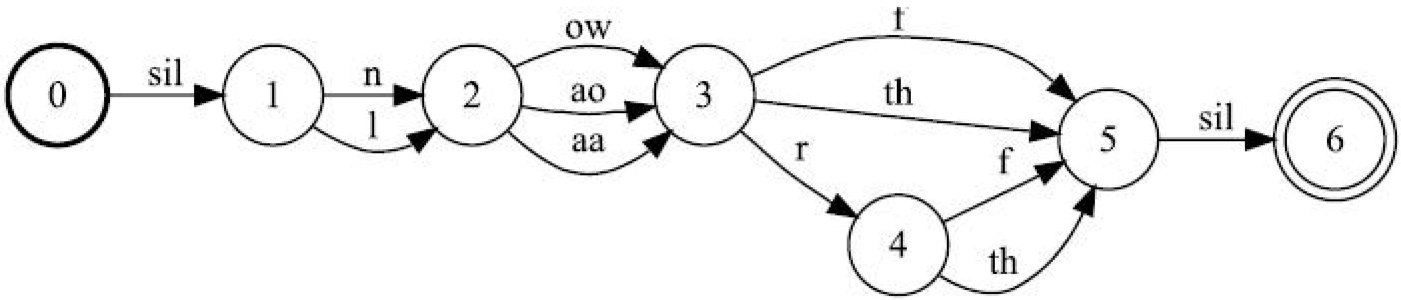


Figure 3: *Extended recognition network of "north"*

# Research ap

- Unsupervised
  - Need larg annotatio

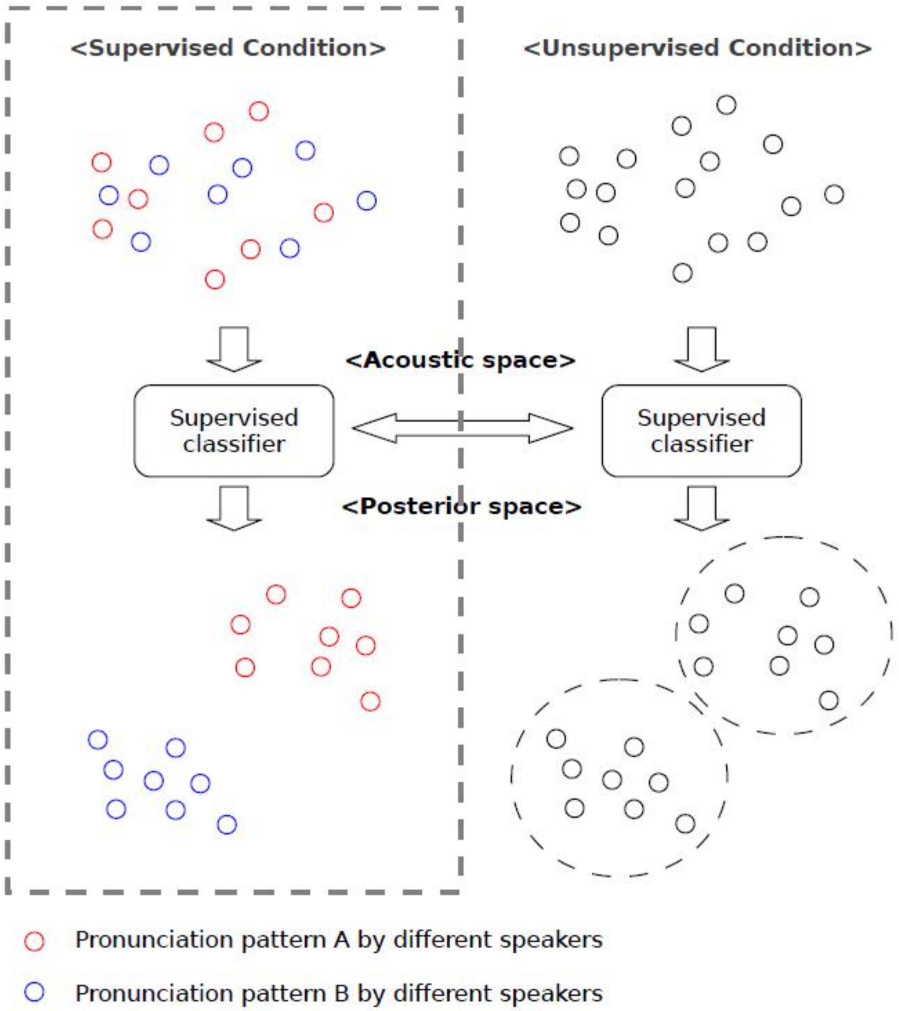


**Fig. 2**. *The effect of mapping from acoustic space to posterior space in supervised and unsupervised learning.*

# Research approaches

- Unsupervised Error Discovery

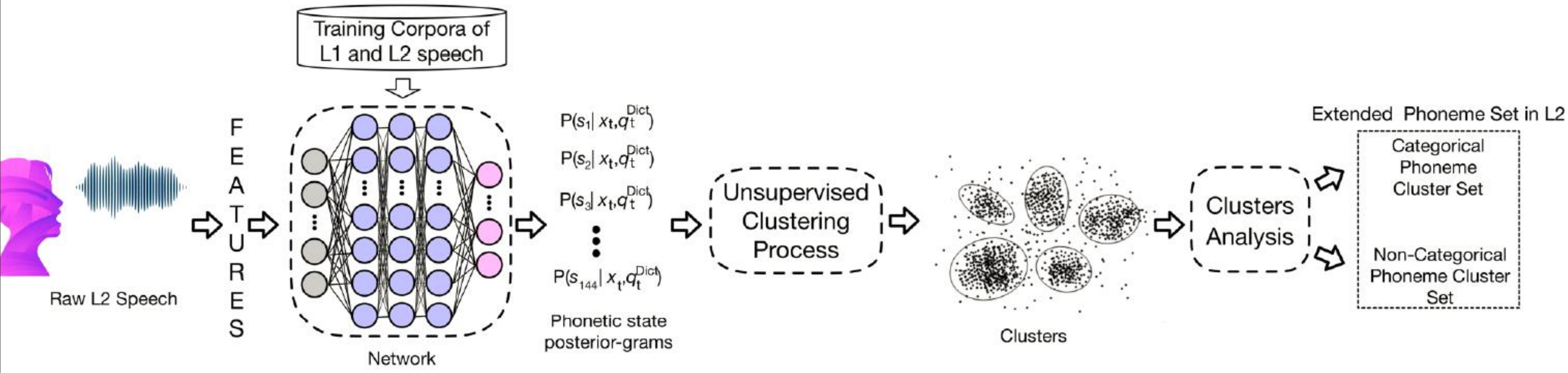
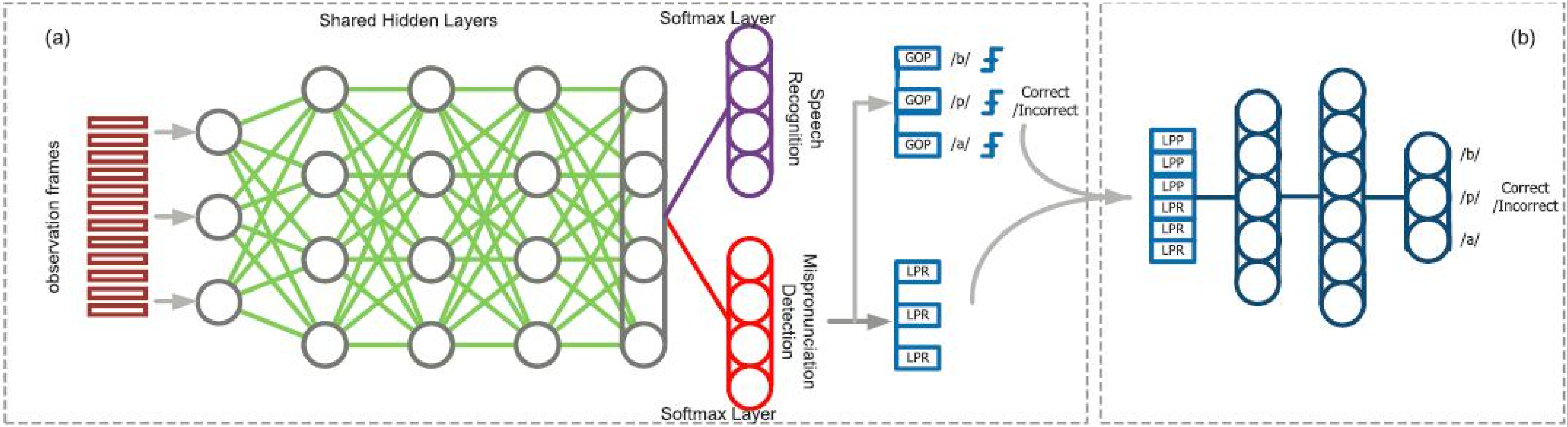

Training Corpora of L1 and L2 speech
Raw L2 Speech
FEATURES
Network
$P(s_1| x_t, q_t^{Dict})$
$P(s_2| x_t, q_t^{Dict})$
$P(s_3| x_t, q_t^{Dict})$
$P(s_{144}| x_t, q_t^{Dict})$
Phonetic state posterior-grams
Unsupervised Clustering Process
Clusters
Clusters Analysis
Extended Phoneme Set in L2
Categorical Phoneme Cluster Set
Non-Categorical Phoneme Cluster Set

**Fig. 2.** The framework of the proposed approach

# Research approaches

- Strategies for Improving Phonetic Error Detection
  - Verification/Rescoring



# Research approaches

- Strategies for Improving Phonetic Error Detection
  - Deep learning
    - DNN-HMM acoustic model better than GMM-HMM baseline
    - Convolutional neural networks were used in to automatically extract features

# Research approaches

- Strategies for Improving Phonetic Error Detection
  - Articulatory or Acoustic Phonetic Knowledge
    - Landmark-based SVM classifiers for detecting possible English pronunciation

# Research approaches

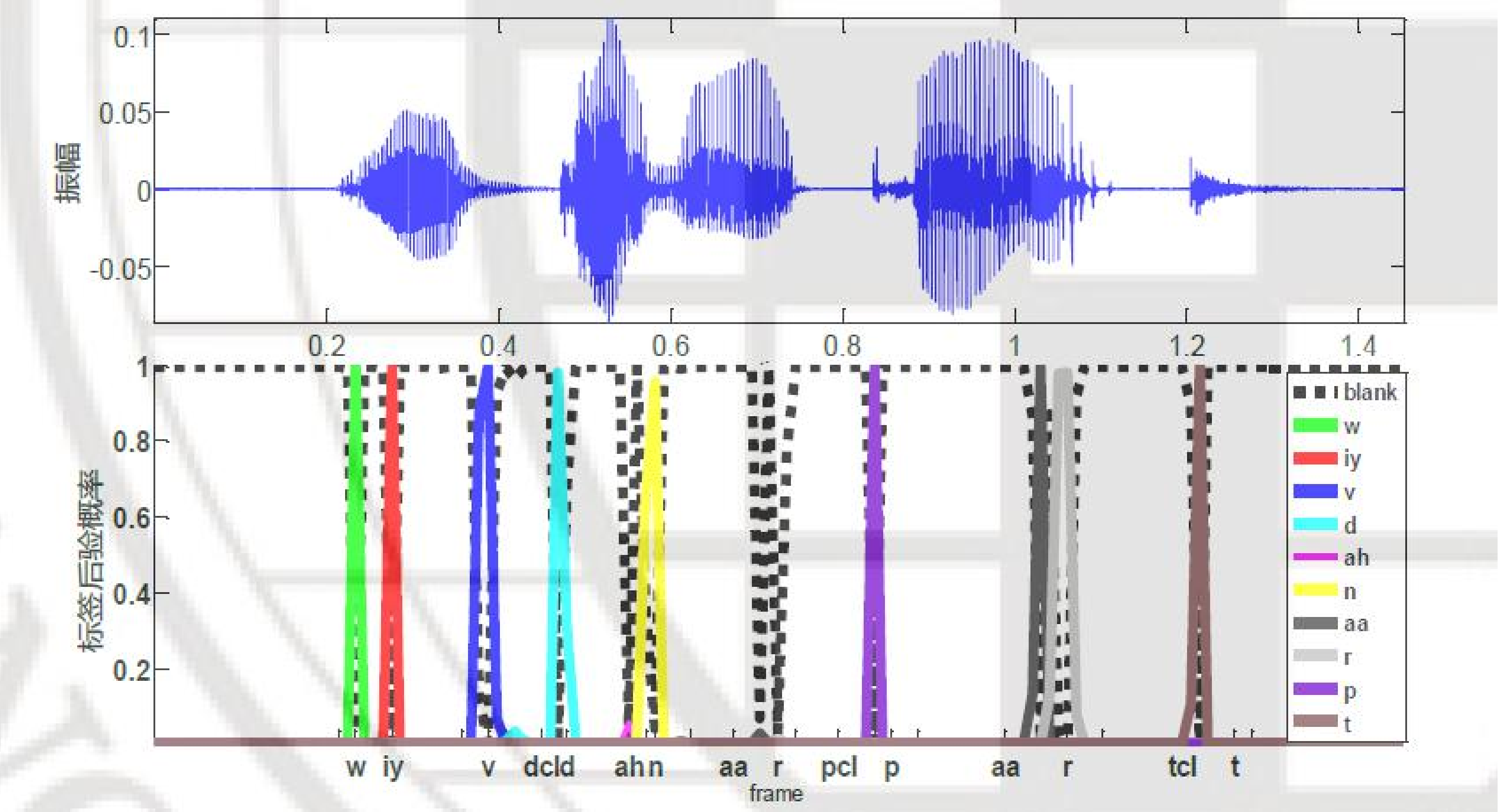


图 15 CTC 的尖峰现象，以“We’ ve done our part”为例

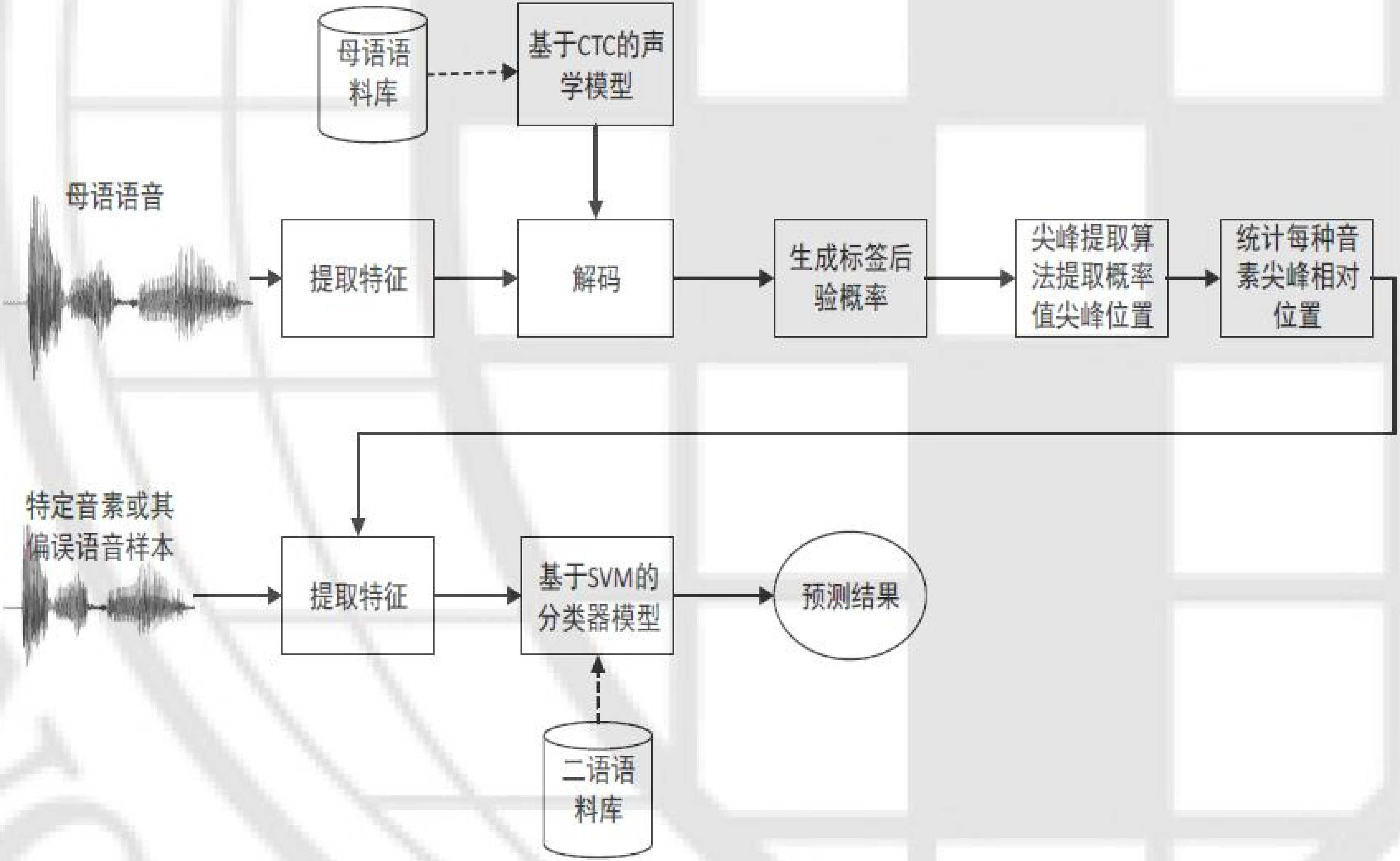


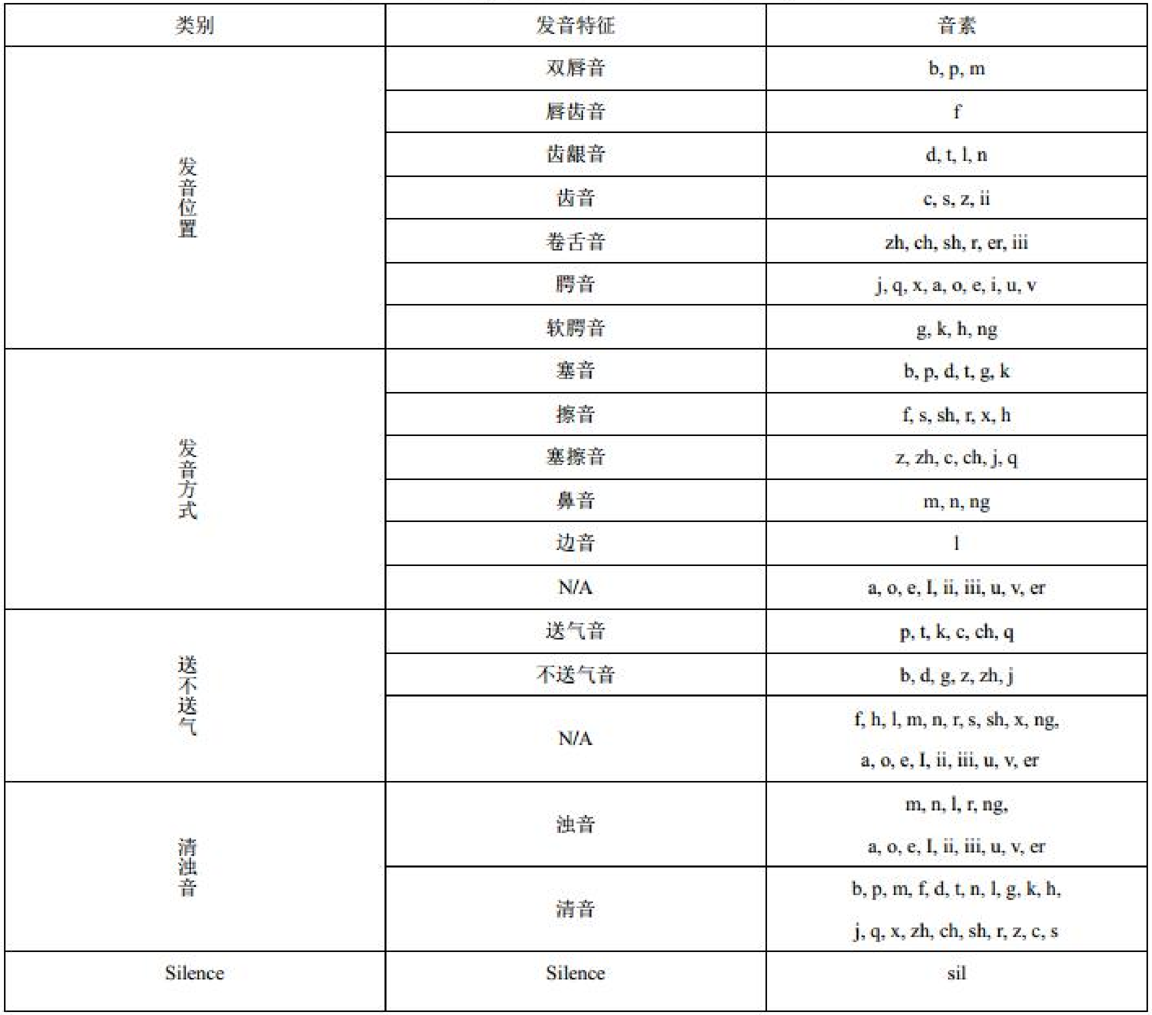
图 17 基于 CTC 的 landmark 检测及其发音偏误检测框架

# Research approaches

- St n

ossible

res(the place



| 类别 | 发音特征 | 音素 |
|---|---|---|
| 发音位置 | 双唇音 | b, p, m |
| | 唇齿音 | f |
| | 齿龈音 | d, t, l, n |
| | 齿音 | c, s, z, ii |
| | 卷舌音 | zh, ch, sh, r, er, iii |
| | 腭音 | j, q, x, a, o, e, i, u, v |
| | 软腭音 | g, k, h, ng |
| 发音方式 | 塞音 | b, p, d, t, g, k |
| | 擦音 | f, s, sh, r, x, h |
| | 塞擦音 | z, zh, c, ch, j, q |
| | 鼻音 | m, n, ng |
| | 边音 | l |
| | N/A | a, o, e, I, ii, iii, u, v, er |
| 送不送气 | 送气音 | p, t, k, c, ch, q |
| | 不送气音 | b, d, g, z, zh, j |
| | N/A | f, h, l, m, n, r, s, sh, x, ng,<br>a, o, e, I, ii, iii, u, v, er |
| 清浊音 | 浊音 | m, n, l, r, ng,<br>a, o, e, I, ii, iii, u, v, er |
| | 清音 | b, p, m, f, d, t, n, l, g, k, h,<br>j, q, x, zh, ch, sh, r, z, c, s |
| Silence | Silence | sil |

# Research approaches

- Detecting Prosodic Errors
  - Lexical Stress
    - Gaussian mixture models perform the best compared to decision trees and neural networks
    - Duration and pitch estimates are the most important features

# Research approaches

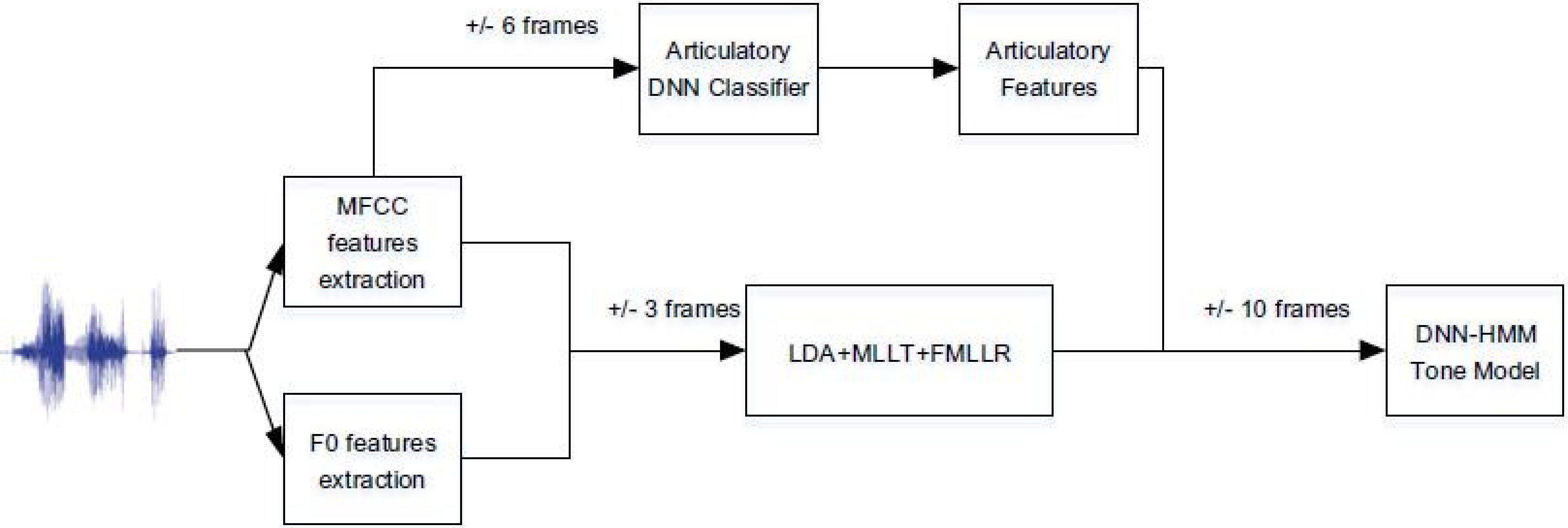
- Detecting Prosodic Errors
  - Lexical Tones
    - Lexical tones are primarily characterized by the pitch contour (e.g., Mandarin), sometimes the pitch height (e.g., Cantonese)
    - ASR framework for Tones Recognition
    - Syllable boundaries+classifier

# Research approaches

- Detecting Prosodic Errors
  - Lexical Tones
    - Segment the F0 contour to tone nucleus
    - Goodness of Tone (GOT)
    - The GOT features were modeled by an SVM classifier
    - Pitch related features could be inferred from a DNN system trained by 40-dimension MFCC features

# Research approaches

- Detecting Prosodic Errors
  - Lexical Tones



# Research approaches

- Automatic Fluency Scoring
  - Cucchiarini et.al found that rate of speech correlates highest with perceptual fluency
  - The number of silent pauses and the rate of articulation

# Challenges And Research Opportunities

- Scarcity of Large-Scale Linguistic Resources
  - Lack of Non-Native Speech Data
    - Substitution phonemic errors by artificially introducing them in a native corpus
  - Lack of Human Annotations
    - Phonetic transcriptions require lots of cost, time, and labor (linguistic expertise)
    - Prosody labeling and fluency scoring can be much more subjective and harder to achieve inter-rater agreement

# Challenges And Research Opportunities

- Common Modeling Assumptions
  - Text dependence
    - The even higher cost of human annotation of datasets if a CAPT system is text-independent
  - Mispronunciations are Categorical
    - Nonnative pronunciations might frequently fall out of the native phonemic or lexical tone categories

# Challenges And Research Opportunities

- Metrics for Evaluation
  - Information retrieval task
    - Precision
    - Recall
  - Mispronunciation detection error
    - False acceptance rate
    - False rejection rate

# References

- Silke M Witt, “Automatic error detection in pronunciation training:Where we are and where we need to go,” in IS ADEPT, 2012, vol. 6
- Xiaojun Qian, Helen M Meng, and Frank K Soong, “The use of dbnhmms for mispronunciation detection and diagnosis in l2 english to support computer-aided pronunciation training.,” in INTERSPEECH,2012, pp. 775–778.
- Wenping Hu, Yao Qian, Frank K Soong, and Yong Wang, “Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers,” Speech Communication, vol. 67, pp. 154–166, 2015.
- Nancy F Chen, Rong Tong, Darren Wee, Peixuan Lee, Bin Ma, and Haizhou Li, “iCALL Corpus: Mandarin Chinese Spoken by Non-Native Speakers of European Descent,” in Sixteenth Annual Conference of the International Speech Communication Association, 2015
- Khiet Truong, Ambra Neri, Catia Cucchiarini, and Helmer Strik, “Automatic pronunciation error detection: an acoustic-phonetic approach,” in InSTIL/ICALL Symposium 2004, 2004
- Yanlu Xie, Mark Hasegawa-Johnson, Leyuan Qu, and Jinsong Zhang, “Landmark of mandarin nasal codas and its application in pronunciation error detection,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 5370–5374.

# References

- Niu C, Zhang J, Yang X, et al. A study on landmark detection based on CTC and its application to pronunciation error detection[C]//Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017. IEEE, 2017: 636-640.
- Lin J , Li W , Gao Y , et al. Improving Mandarin Tone Recognition Based on DNN by Combining Acoustic and Articulatory Features Using Extended Recognition Networks[J]. Journal of Signal Processing Systems, 2018.

谢谢

请大家批评指正