# Exploring the Use of Deep Learning Techniques for Mandarin Mispronunciation Detection

# 深層學習技術在華語錯誤發音偵測之研究

Berlin Chen (陳柏琳)
Professor, Department of Computer Science & Information Engineering
National Taiwan Normal University
2016/4/20

SMIL

# Outline

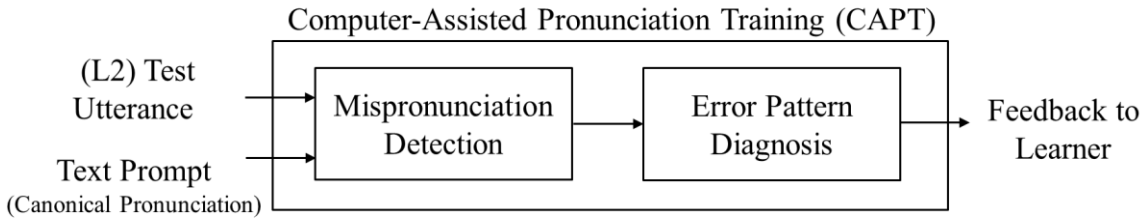
- Introduction
- Typical Framework for MD
- Leveraging Deep Leaning Technology for MD
- Maximum Performance Criterion Training for Acoustic Models and Decision Functions
- Experimental Results
- Conclusion and Outlook

Yao-Chi Hsu, Ming-Han Yang, Hsiao-Tsung Hung, Berlin Chen, "Mispronunciation detection leveraging maximum performance criterion training of acoustic models and decision functions," the 17th Annual Conference of the International Speech Communication Association (Interspeech 2016), San Francisco, USA, September 8-12, 2016.

SMIL

# Introduction

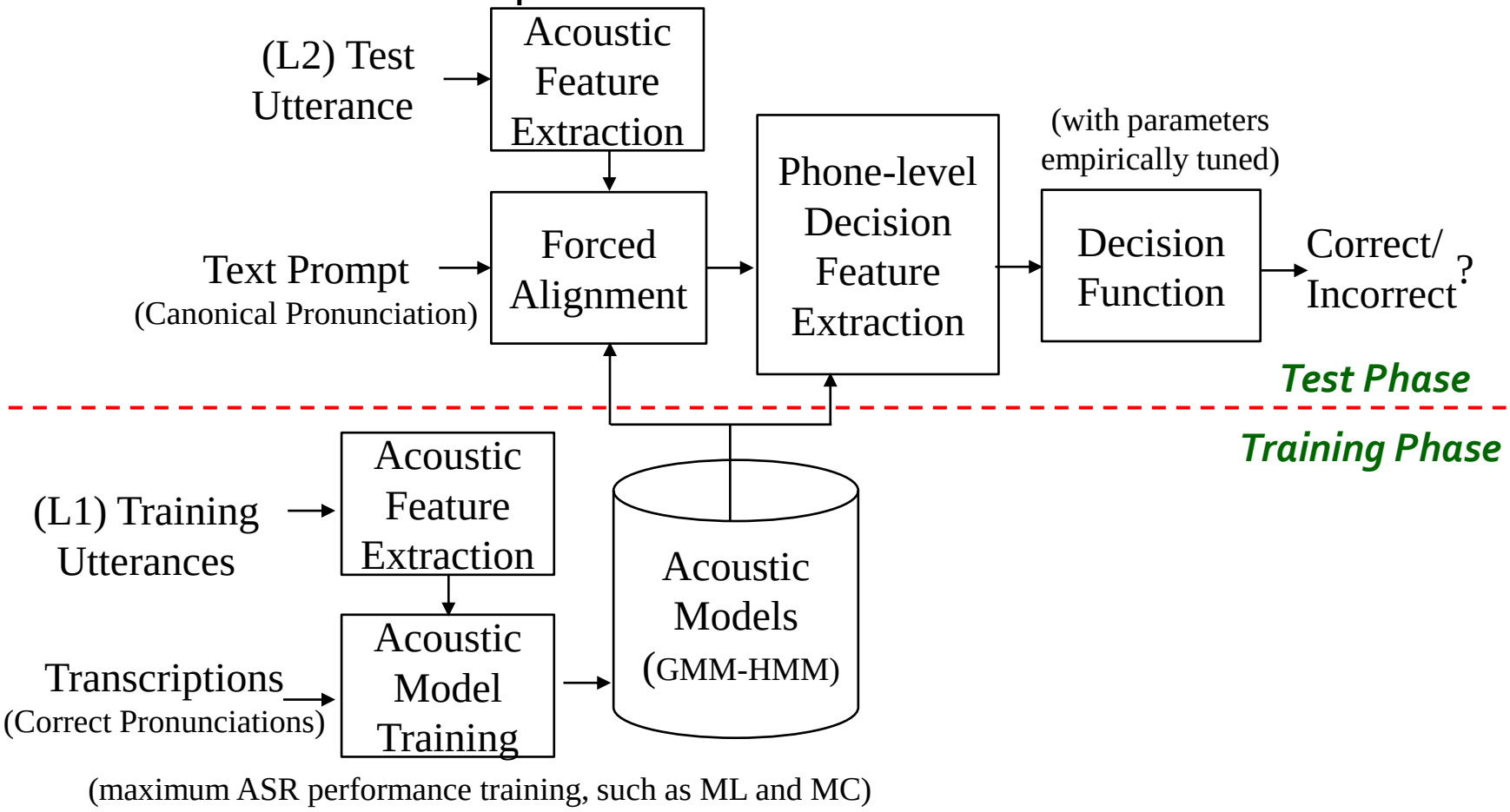
- Computer assisted pronunciation training (CAPT) has attracted increasing research interest recently, partly due to the rapid progress of automatic speech recognition (ASR) technology
  - Deep Learning + Increasing Computational Power + Big Data + …



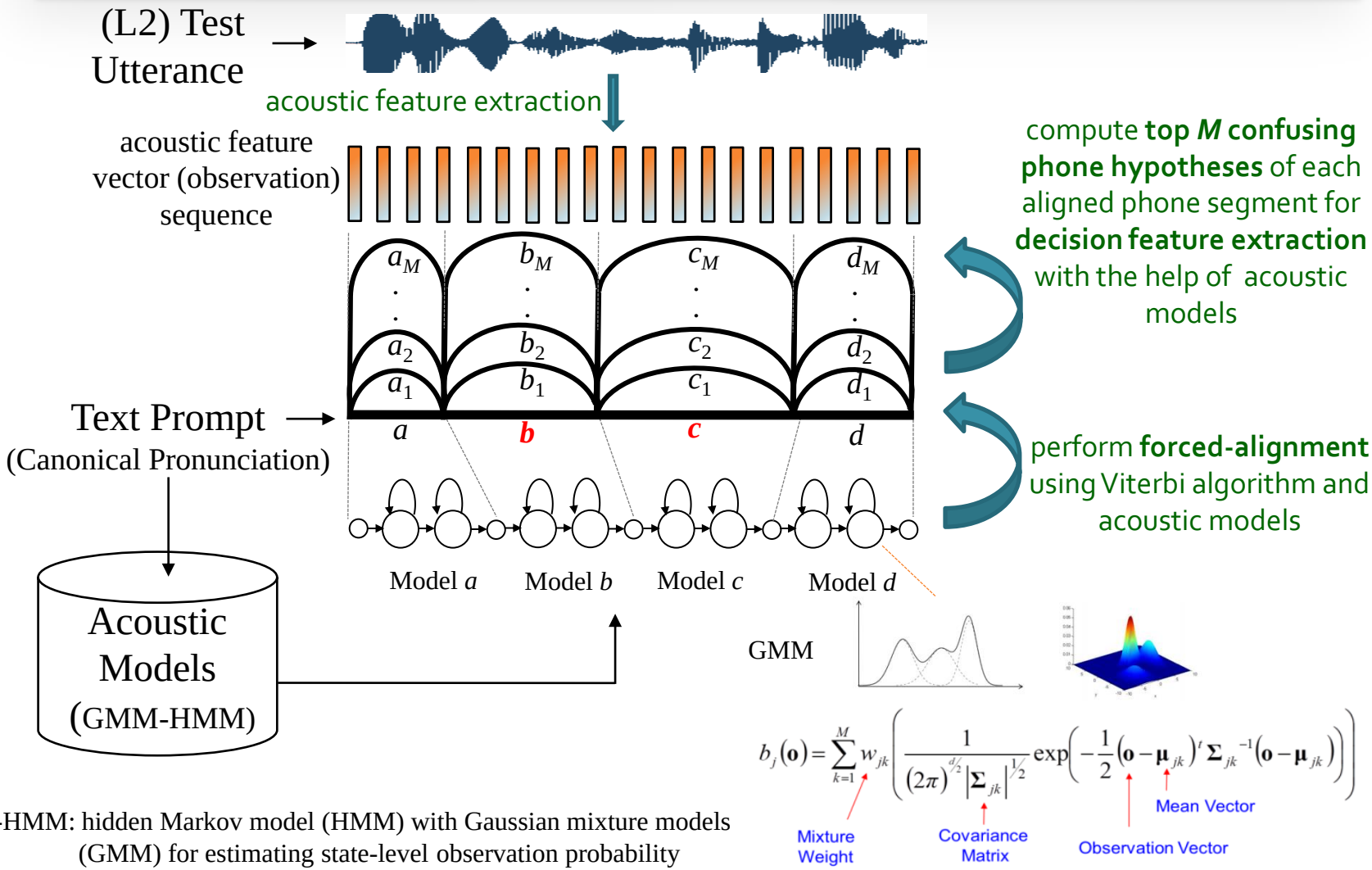

- Mispronunciation detection (MD) is an essential module in a CAPT system
  - Assist second-language (L2) learners to pinpoint incorrect pronunciations in a given utterance in order to improve their spoken proficiency
  - E.g., phone-level or word-level substitution errors, insertion errors, deletion errors, among others

# Technical Framework for MD

- Schematic diagram of **a conventional (mainstream) framework** for mispronunciation detection



# Forced Alignment & Generating Competing Phone Hypotheses (in the Test Phase)




GMM-HMM: hidden Markov model (HMM) with Gaussian mixture models (GMM) for estimating state-level observation probability

# Phone-level Decision Feature Extraction

- Adopt the commonly-used **goodness of pronunciation** (**GOP**) measure for decision feature extraction, based on the phone-level posterior probabilities computed with forced alignment and acoustic models

$$\mathrm{GOP}(u,n) = \frac{1}{T_{u,n}} \log P(q_{u,n} \mid \mathbf{O}_{u,n})$$

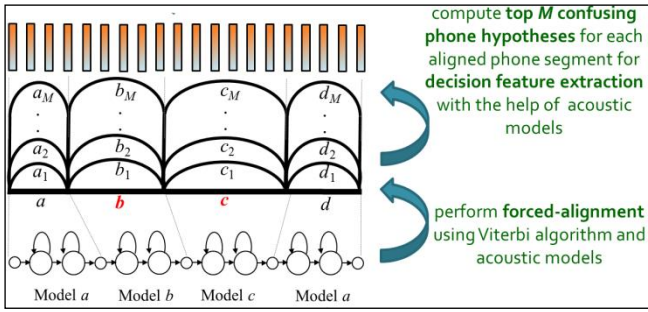
posterior probability

$$\approx \frac{1}{T_{u,n}} \log \frac{P(\mathbf{O}_{u,n} \mid q_{u,n})}{\sum_{\tilde{q} \in \{\mathrm{Top}\ M\}} P(\mathbf{O}_{u,n} \mid \tilde{q})}$$

log likelihood ratio

or

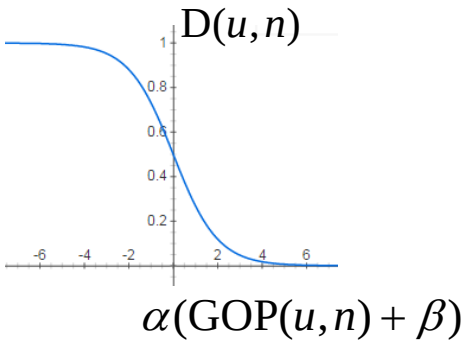
$$\mathrm{GOP}(u,n) \approx \frac{1}{T_{u,n}} \log \frac{P(\mathbf{O}_{u,n} \mid q_{u,n})}{\max_{\tilde{q} \in \{Top\ \mathrm{M}\}} P(\mathbf{O}_{u,n} \mid \tilde{q})}$$



# Phone-level Decision Functions

- As to the decision function, we can adopt the **logistic sigmoid function** for our purpose

$$\mathrm{D}(u,n) = \frac{1}{1+\exp\left[\alpha(\mathrm{GOP}(u,n)+\beta)\right]}$$




- Take the GOP score as the input and output a decision score, ranging between 0 and 1
- $\mathrm{D}(u,n) \geq \tau$ implies the occurrence of mispronunciation for phone $q_{u,n}$
  - The higher the decision score, $\mathrm{D}(u,n)$ , the more likely the phone $q_{u,n}$ is mispronounced
- The parameters $\alpha, \beta$ and the threshold $\tau$ are empirically tuned in practice (one size fits all: all phones share the same set of parameters/threshold)

SMIL

# Our Research Contributions for MD (1/2)

1. We explore recent advances in **deep learning** (especially **deep neural networks**, **DNN**) to achieve better speech feature extraction and acoustic modeling

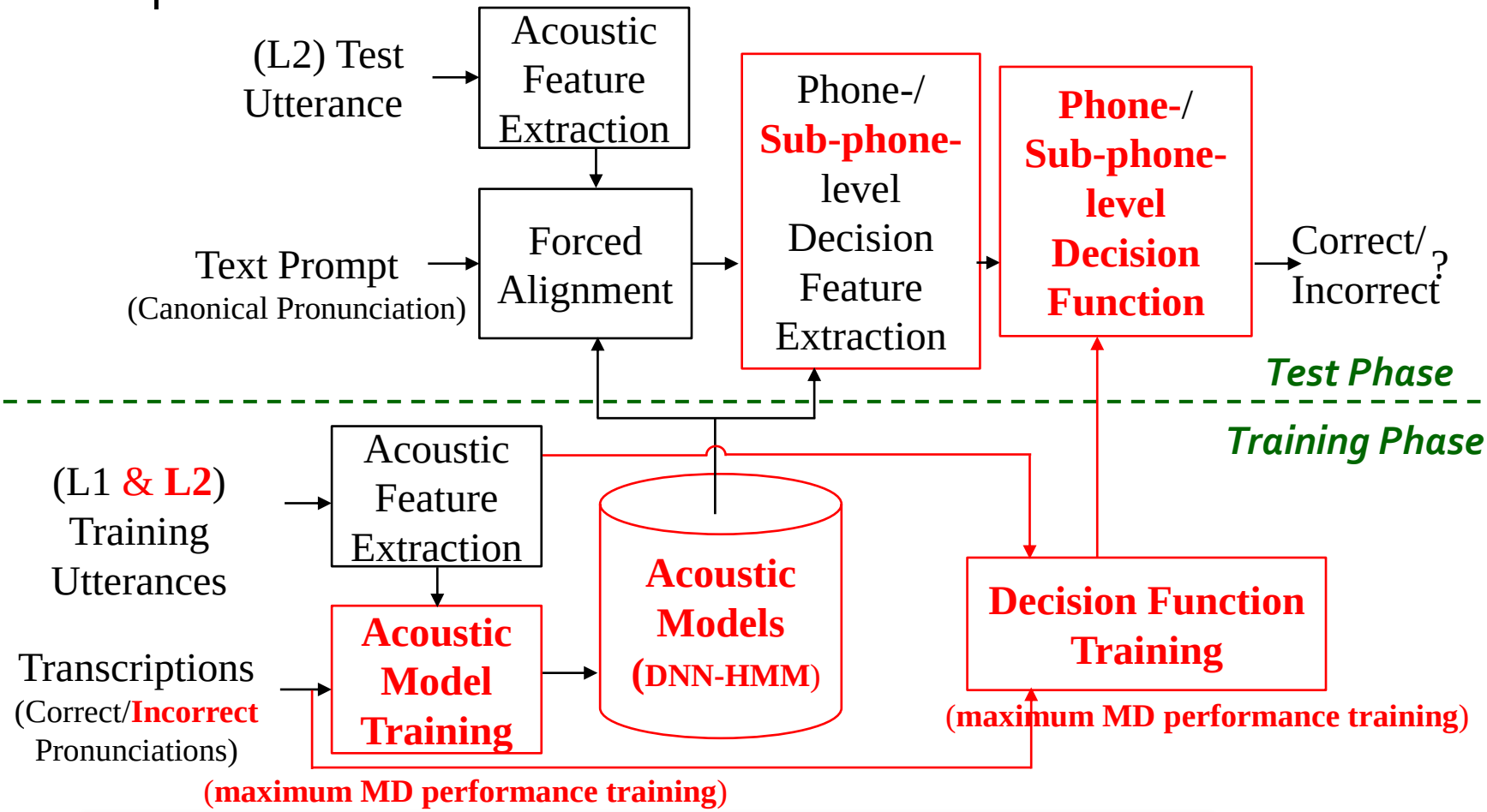
2. An effective learning approach is proposed, which estimates the DNN-based acoustic models by optimizing an objective directly linked to the ultimate evaluation metric of mispronunciation detection

3. Decision functions of different levels of granularity, with either phone- or sub-phone(senone)-dependent parameterization, are also explored for mispronunciation detection

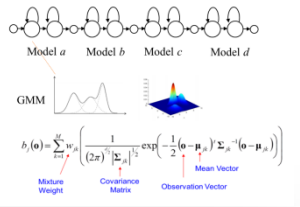
SMIL

# Our Research Contributions for MD (2/2)

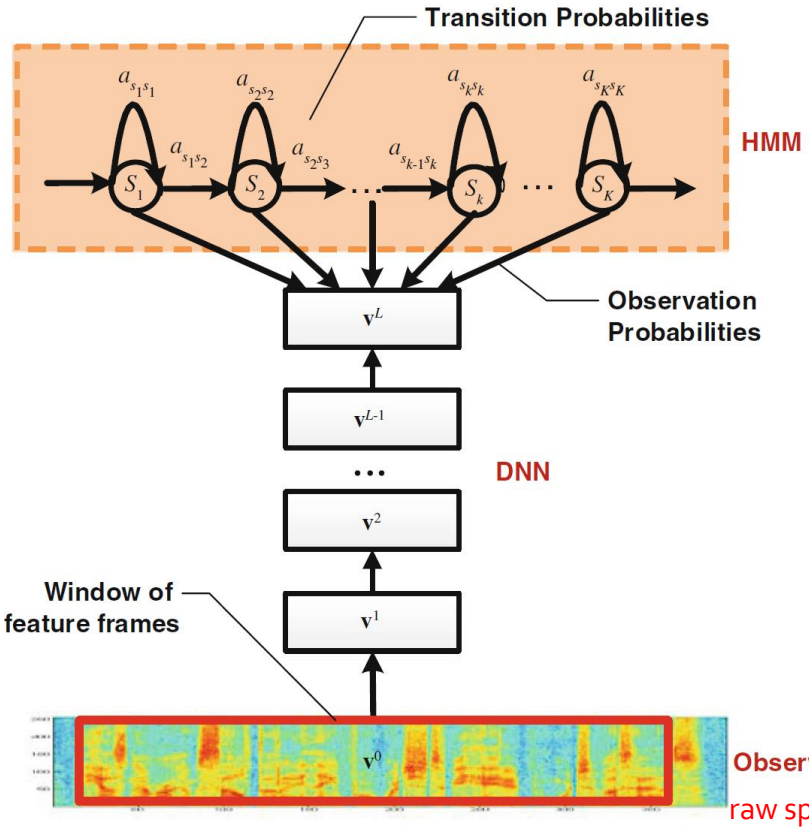
- Schematic diagram of **our proposed approach** to mispronunciation detection



# 1. Deep Learning for Acoustic Modeling




- We leverage various state-of-the-art deep neural network (**DNN**) architectures (in place of **GMM** ) for modeling the state emission probabilities in HMM (denoted by DNN-HMM)




deeper layers,
longer features &
wider temporal contexts

$$b_{s_i}(\mathbf{o}) = p(\mathbf{o} \mid s_i) = \frac{P_{\text{DNN}}(s_i \mid \mathbf{o})\, p(\mathbf{o})}{P_{\text{ML}}(s_i)} \propto \frac{P_{\text{DNN}}(s_i \mid \mathbf{o})}{P_{\text{ML}}(s_i)}$$

$$P_{\text{DNN}}(s_i \mid \mathbf{o}) = v_i^L = \text{softmax}_i\left(\mathbf{z}^L\right) = \frac{e^{z_i^L}}{\sum_j e^{z_j^L}}$$

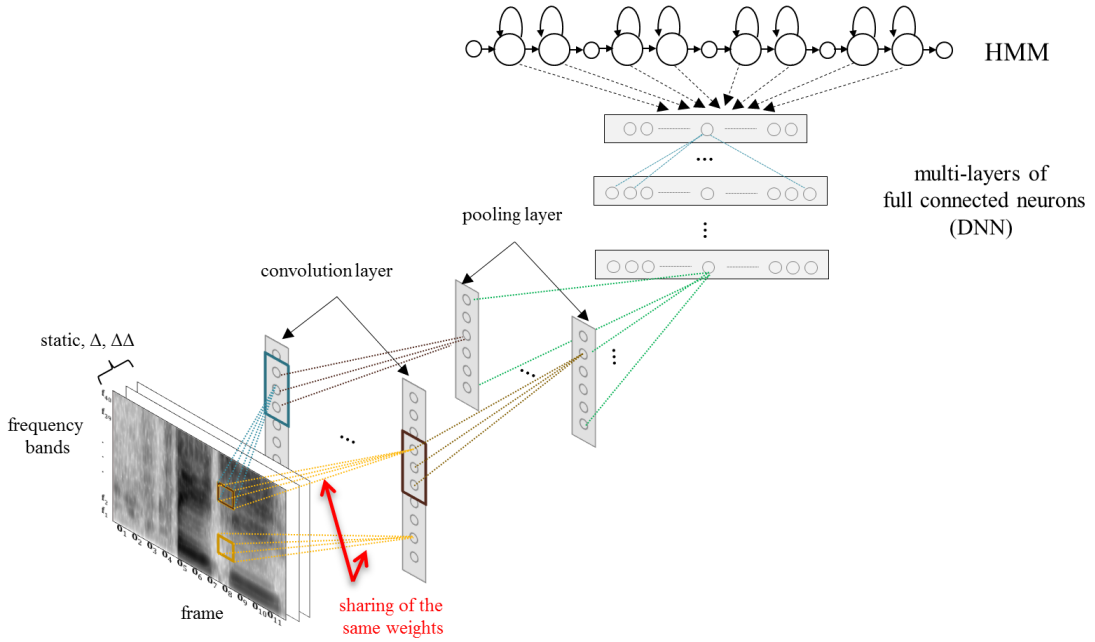
$$\mathbf{v}^\ell = f\left(\mathbf{z}^\ell\right) = f\left(\mathbf{W}^\ell \mathbf{v}^{\ell-1} + \mathbf{b}^\ell\right), \text{ for } 0 < \ell < L$$

$f(\cdot)$ : sigmoid, hyperbolic, or rectified linear unit (ReLU) functions

Model parameters of DNN can be estimated with the **error back-propagation algorithm** and **stochastic gradient decent** (SGD).

# CNN for Acoustic Modeling in MD

- Alternatively, we also explore to use the convolutional neural networks (CNN) to replace GMM for predicting the state-level likelihoods of acoustic feature vectors
  - Schematic Depiction of Using Convolutional Neural Networks (CNN) for acoustic modeling (i.e., CNN-HMM)



# 2. Maximum Performance Criterion Training for MD

- Instead of training the acoustic models with criteria that maximize the ASR performance, we attempt to train the acoustic models with an objective function that directly maximizes the performance of MD
  - For example, the **maximum F1-score criterion** (**MFC**)

$$\Xi(\boldsymbol{\theta}) = \frac{2C_{\mathrm{D}\cap\mathrm{H}}}{C_{\mathrm{D}} + C_{\mathrm{H}}} = \frac{2 \cdot \sum_{u=1}^{U} \sum_{n=1}^{N_u} \mathrm{I}(\mathrm{D}(u,n)) \cdot \mathrm{H}(u,n)}{[\sum_{u=1}^{U} \sum_{n=1}^{N_u} \mathrm{I}(\mathrm{D}(u,n))] + C_{\mathrm{H}}}$$

$$\approx \frac{2 \cdot \sum_{u=1}^{U} \sum_{n=1}^{N_u} \mathrm{D}(u,n) \cdot \mathrm{H}(u,n)}{[\sum_{u=1}^{U} \sum_{n=1}^{N_u} \mathrm{D}(u,n)) + C_{\mathrm{H}}}$$

    - Where $\boldsymbol{\theta}$ denotes the set of **parameters** of both **the DNN-HMM based acoustic models** and **the decision function**
    - $C_{\mathrm{D}\cap\mathrm{H}}$ is the total number of phone segments in the training set that are identified as being mispronounced simultaneously by both the current mispronunciation detection module and the majority vote of human assessors
    - Optimized by stochastic gradient ascent algorithm + chain rule for differentiation

# Appendix: F1-Score for Performance Evaluation

- The default evaluation metric for **mispronunciation detection** employed in this work is the F1-score, which is a harmonic mean of precision and recall

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{C_{D \cap H}}{C_D + C_H}$$

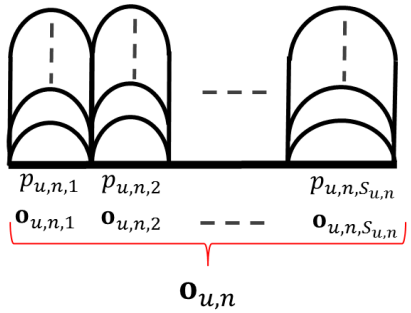
$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} = \frac{C_{D \cap H}}{C_D}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} = \frac{C_{D \cap H}}{C_H}$$

# 3. Sub-phone-level Decision Functions

- We explore to obtain a finer-grained inspection of the pronunciation quality of a phone segment $\mathbf{O}_{u,n}$ by using sub-phone-level decision functions

$$\mathrm{D}(u,n) = \frac{1}{S_{u,n}} \sum_{i=1}^{S_{u,n}} \tilde{\mathrm{D}}(u,n,i)$$




top *M* confusing sub-phone hypotheses for each sub-phone segment

the corresponding canonical sub-phone models

sub-phone segments

phone segment

- $\tilde{\mathrm{D}}(u,n,i)$ is the sub-phone-level decision function
- $S_{u,n}$ is the total number of sub-phone segments $\mathbf{O}_{u,n,i}$ corresponding to the phone segment $\mathbf{O}_{u,n}$

- The above equation represents an ensemble of the output scores of all sub-phone-level decision functions for $\mathbf{O}_{u,n}$
- Each sub-phone-level decision function can be optimized with the proposed MFC training criterion & sub-phone-dependent parameterization

# Experimental Corpus (1/2)

- The dataset employed in this study is a Mandarin annotated spoken (MAS) corpus compiled by the **Center of Learning Technology for Chinese**, **National Taiwan Normal University**, between 2012 and 2014 [1]

| | | Duration (hours) | # Speakers | # Phone Tokens | # Errors |
|---|---|---|---|---|---|
| Training Set | L1 | 6.68 | 44 | 73,074 | NA |
| | L2 | 15.79 | 63 | 118,754 | 26,434 |
| Development Set | L1 | 1.40 | 10 | 14,216 | NA |
| | L2 | 1.46 | 6 | 11,214 | 2,699 |
| Test Set | L1 | 3.20 | 26 | 32,568 | NA |
| | L2 | 7.49 | 44 | 55,190 | 14,247 |

  - Utterances of L2 learners may contain mispronunciations, each of which was carefully cross-checked by 2 to 4 human assessors

1. Y. Hsiung, B. Chen, and Y. Sung, "Development of Mandarin annotated spoken corpus (MAS Corpus) and the learner corpus analysis," in Proc. WoALF, 2014.

# Experimental Corpus (2/2)

- The corpus was split into three subsets: training set, development set and test set
- All these subsets are composed of speech utterances (containing one to several syllables) pronounced by native speakers (L1) and L2 learners
  - Monosyllables:
    - 剛 (gang1) 、 王 (wang2) 、 咬 (yao3) 、 練 (lian4) …
  - Disyllables:
    - 飛機 (fei1 ji1) 、 炒麵 (chao3 mian4) …
  - Polysyllables：
    - 王冕 自此 在 秦家 放牛 (wang2 mian3 zi4 ci3 zai4 qin2 jia1 fang4 niu2)

# Baseline ASR Performance

- Compare GMM-HMM with DNN-HMM for acoustic modeling in terms of ASR Performance (on the L1 portion of the test set)
  - Free-syllable decoding without language model constraints
  - The lower the SER and PER, the better the ASR performance

| | Syllable Error Rate, SER (%) | Phone Error Rate, PER (%) |
|---|---|---|
| GMM-HMM | 50.9 | 34.3 |
| DNN(A)-HMM | 41.2 | 27.7 |
| DNN(B)-HMM | 40.1 | 27.0 |
| DNN(C)-HMM | 40.7 | 27.2 |
| DNN(B)-HMM+sMBR | 37.9 | 24.9 |

  - Different model structures for DNN-HMM

| | # Layers | # Neurons per Layer |
|---|---|---|
| DNN(A)-HMM | 4 | 1,024 |
| DNN(B)-HMM | 4 | 2,048 |
| DNN(C)-HMM | 6 | 1,024 |

  - DNN-HMM shows significant performance gains over GMM-HMM

Since the ASR results on CNN-HMM are not as significantly improved as DNN-HMM, we omit the experimental results with CNN-HMM hereafter.

# Performance of Mispronunciation Detection (1/4)

- Mispronunciation detection results achieved by using either the phone- or the sub-phone(senone)-level decision function and with or without the proposed MFC training

| | Recall | Precision | F1 Score |
|---|---|---|---|
| Phone-level | 0.681 | 0.537 | 0.600 |
| Senone-level | 0.675 | 0.545 | 0.603 |
| +MFC (Both) | 0.696 | 0.626 | 0.659 |
| +MFC (AM) | 0.697 | 0.621 | 0.657 |
| +MFC (DF) | 0.688 | 0.581 | 0.630 |

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{C_{D \cap H}}{C_D + C_H}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} = \frac{C_{D \cap H}}{C_D}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} = \frac{C_{D \cap H}}{C_H}$$

  - The acoustic models are DNN(B)-HMM trained with minimum cross-entropy (MC) criterion
  - MFC (AM): the MFC training was applied on the acoustic models
  - MFC (DF): the MFC training was applied on the decision functions for all sub-phone units

# Performance of Mispronunciation Detection (2/4)

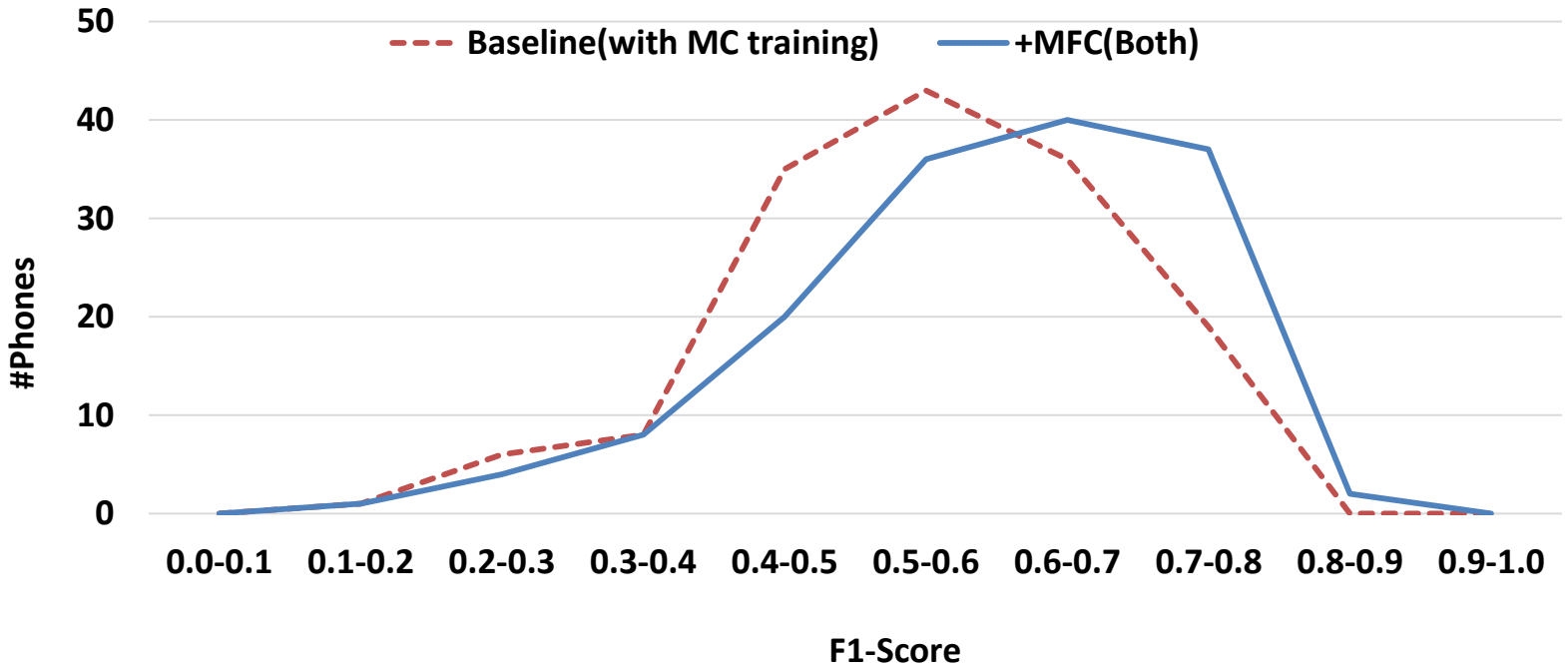
- Acoustic models were first pre-trianed with a conventional **ASR-oriented** discriminative training criterion (i.e., **sMBR**), and subsequently trained with our proposed **MD-oriented** training criterion (i.e., **MFC**)

| | Recall | Precision | F1 Score |
|---|---|---|---|
| Phone-level | 0.671 | 0.551 | 0.605 |
| Senone-level | 0.652 | 0.555 | 0.599 |
| +MFC (Both) | 0.743 | 0.587 | 0.656 |
| +MFC (AM) | 0.738 | 0.586 | 0.653 |
| +MFC (DF) | 0.698 | 0.570 | 0.627 |

  - Even though sMBR can considerably improve the ASR performance in terms of SER and PER, it does not provide any additional gain for mispronunciation detection
    - When employing either the MC-estimated acoustic models or the acoustic models further trained with the MFC criterion

# Performance of Mispronunciation Detection (3/4)

- Plots of the F1-Score Distributions, before and after the MFC training (for mispronounced phone segments)




  - An obvious shift of the distribution toward the right (i.e., the direction of higher F1-Scores)

# Performance of Mispronunciation Detection (4/4)

- Graphical inspection of the performance of different MD methods

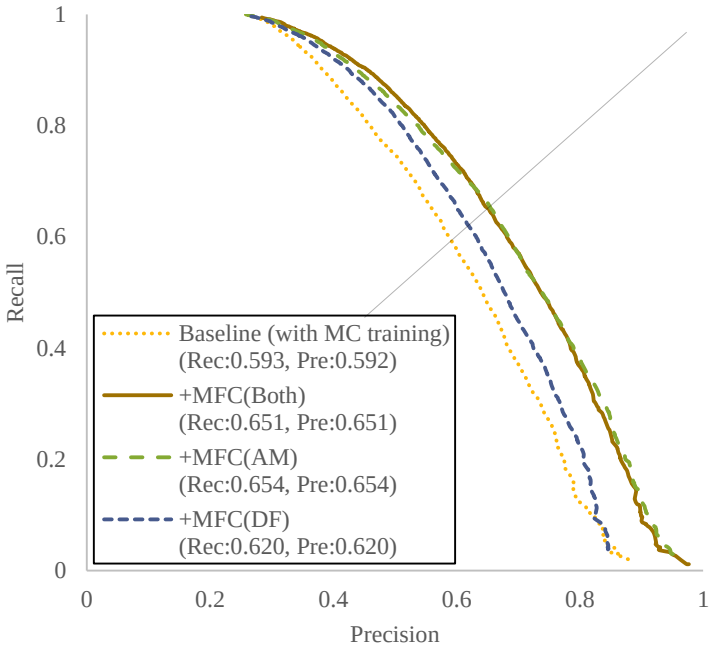



Figure 1: *Recall-precision curves for different training settings shown in Table 4 (with the senone-level decision function).*

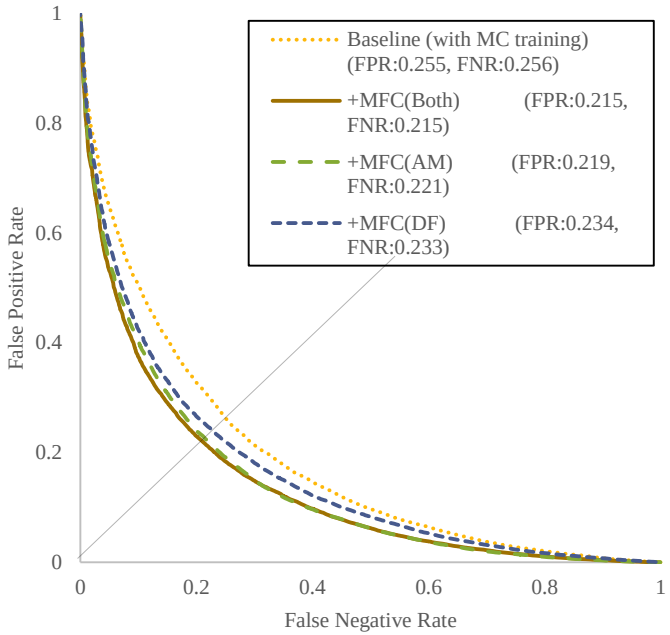



Figure 2: *ROC curves for different training settings shown in Table 4 (with the senone-level decision function).*

# Performance of Correct Pronunciation Detection (1/2)

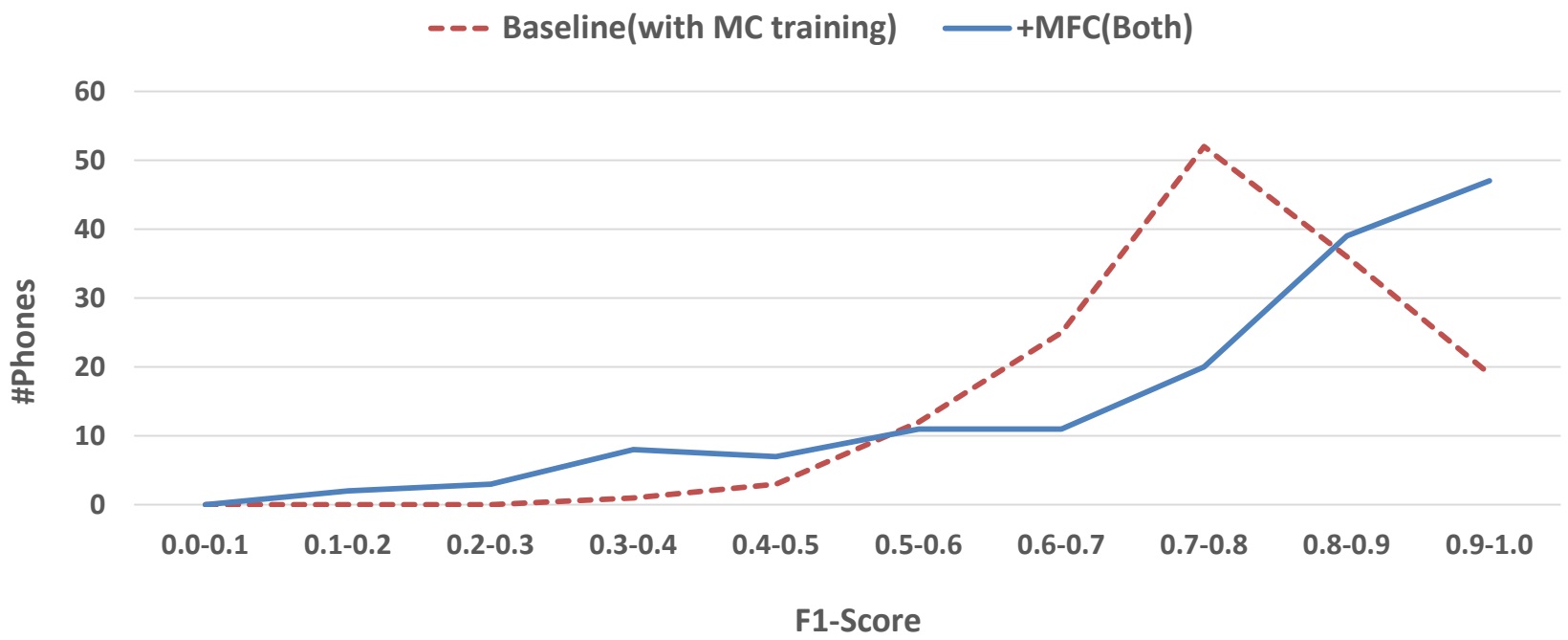
- Correct pronunciation detection results achieved by using either the phone- or the sub-phone(senone)-level decision function and with or without the proposed MFC training

| | Recall | Precision | F1 Score |
|---|---|---|---|
| Phone-level | 0.795 | 0.878 | 0.834 |
| Senone-level | 0.804 | 0.877 | 0.839 |
| +MFC (Both) | 0.828 | 0.884 | 0.855 |
| +MFC (AM) | 0.852 | 0.890 | 0.871 |
| +MFC (DF) | 0.840 | 0.896 | 0.867 |

  - The recall, precision and F1 scores for detecting the correctly pronounced phone segments can also be considerably improved

# Performance of Correct Pronunciation Detection (2/2)

- Plots of the F1-Score Distributions, before and after the MFC training (for detecting the correctly pronounced phone segments of L2 learners)



# Prototype System

- This prototype system is built and maintained by Prof. Yao-Ting Sung's research group


http://smartpinyin.net/SmartPinyinDesign/Pages/Home.aspx

# Conclusion & Outlook

- We have explored an effective maximum performance criterion training (i.e., **MFC**) approach for estimating the deep neural network based acoustic models and the logistic sigmoid based decision functions involved in mispronunciation detection
  - Both phone- and sub-phone-level decision functions were also investigated
  - Empirical evidence confirms the utility of the proposed approach
- We plan to collectively integrate more acoustic & prosodic features, and other different kinds of speaking-style information cues (manners of articulation), into the mispronunciation detection process

SMIL

# *Thank You!*