# INTEGRATING ARTICULATORY FEATURES INTO ACOUSTIC-PHONEMIC MODEL FOR MISPRONUNCIATION DETECTION AND DIAGNOSIS IN L2 ENGLISH SPEECH

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

[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

{msg16, lirn15}@mails.tsinghua.edu.cn, {zywu, xuli, wuxx, hmmeng}@se.cuhk.edu.hk

## ABSTRACT

This paper proposes novel approaches to mispronunciation detection and diagnosis (MDD) on second-language (L2) learners' speech with articulatory features. Here, articulatory features are the positions of articulators when pronouncing phonemes and reflect the pronunciation mechanisms of each phoneme. The use of articulatory features in MDD is helpful in distinguishing phonemes. Three models with articulatory features are proposed based on acoustic-phonemic model (APM): 1) articulatory-acoustic-phonemic model (AAPM) that embeds articulatory features directly into input features; 2) AAPM with feature representation (R-AAPM) to re-represent original input features with articulatory features; and 3) articulatory multi-task acoustic-phonemic model (A-MT-APM) where phoneme recognizer and articulatory feature classifiers are trained simultaneously in multi-task manner. Compared with baseline phoneme-based APM, proposed approaches perform better in mispronunciation detection and diagnosis measured with Precision, Recall and F1-Measure metrics. Specifically, the A-MT-APM approach gains 5.6% and 7.0% improvement in F1-Measure and diagnostic accuracy respectively. The contributions include: 1) introducing the articulatory features to MDD in deep learning framework; 2) investigating several model architectures for better exploiting articulatory features.

***Index Terms***— Computer-aided pronunciation training, mispronunciation detection and diagnosis, articulatory features, acoustic-phonemic model, multi-task learning

## 1. INTRODUCTION

Computer-aided pronunciation training (CAPT) systems provide second-language (L2) learners with self-learning opportunities and thus have drawn increasing research interests. As the core of CAPT, mispronunciation detection and diagnosis (MDD) aims to detect mispronunciations from learners' L2 speech and further give diagnostic feedbacks [1]. Nowadays MDD techniques can be grouped into two categories: 1) pronunciation scoring based approaches that use different confidence measures to evaluate pronunciations and give evaluation scores as feedback [2]-[6]; 2) automatic speech recognition (ASR) derived approaches by treating MDD as a specialized ASR problem and trying to detect mispronunciations and give feedbacks about specific errors such as phoneme substitutions, deletions and insertions based on recognition results [7]-[13]. Goodness of pronunciation (GOP) [3] is the most representative scoring method, where GOP is obtained from target phoneme's posterior probability computed by the acoustic models. Although scoring methods can judge whether the pronunciation is correct or not by setting thresholds, they are limited in generating specific diagnostic feedbacks. As a specialized ASR derived MDD method, extended recognition networks (ERN) have achieved commendable performance [10]-[12], which incorporate manually designed or data-driven derived phonological rules and design the possible phoneme paths in a word, including both canonical phonetic path and common mispronunciation paths. However, these techniques have limited performance in diagnosing mispronunciation patterns uncovered by the phonological rules. In dealing with the problem, acoustic-phonemic model (APM) has been proposed [13], which employs deep neural networks (DNN) to map the acoustic information and phonetic context information into phonetic posterior probabilities. Because of the incorporation of canonical phonetic context information, APM can achieve better performance in MDD.

Comparing to conventional ASR task, the challenge in MDD is more complex: pronunciations from L2-learners are nonstandard and unordered so it is hard to get a reliable diagnosis based on recognition results. With the development of ASR techniques [14] and deep learning technologies [15]-[16], MDD has achieved significant improvements on accuracy detection, but further improvement on accuracy diagnosis is still needed. In recent researches, articulatory features are widely employed to describe the pronunciation mechanisms and positions of articulators when pronouncing [17], which can indicate the exclusive information of each phoneme. Some previous work has already employed articulatory features to boost ASR and MDD with decision tree [18] and Gaussian mixture model-hidden Markov model (GMM-HMM) [19]-[22] etc.

This work 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. The research was conducted while the first author was an intern at CUHK.

978-1-5386-1737-3/18/$31.00 ©2018 IEEE

**Table 1.** The articulatory feature space in this work [22]

| *stream* | *classes* | *cardinality* |
|---|---|---|
| *jaw* | 0: Nearly Closed, 1: Neutral, 2: Slightly Lowered, 3: Lowered | 4 |
| *lip separation* | 0: Closed, 1: Slightly Apart, 2: Apart, 3: Wide Apart | 4 |
| *lip rounding* | 0: Rounded, 1: Slightly Rounded, 2: Neutral, 3: Spread | 4 |
| *tongue frontness* | 0: Back, 1: Slightly Back, 2: Neutral, 3: Slightly Front, 4: Front | 5 |
| *tongue height* | 0: Low, 1: Mid, 2: Mid-High, 3: High | 4 |
| *tongue tip* | 0: Low, 1: Neutral, 2: Dental, 3: Nearly Alveolar, 4: Alveolar | 5 |
| *velum* | 0: Closed, 1: Open | 2 |
| *voicing* | 0: Unvoiced, 1: Voiced | 2 |

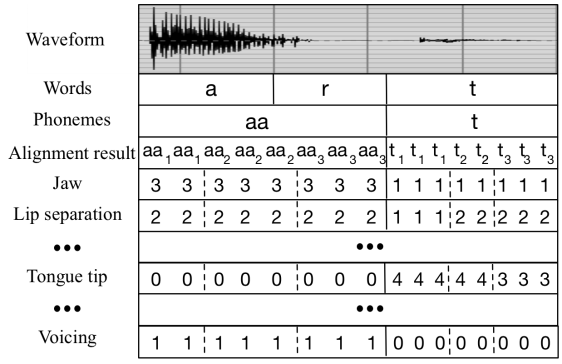


**Fig. 1.** Illustration of how to map articulatory features to corresponding speech frames. (phoneme$_i$ is the $i$th HMM state of this phoneme)

However, due to the limitations of models, information from articulatory features isn't utilized effectively. To further improve the performance of MDD, especially for diagnosis, in this paper we introduce the articulatory features to MDD in deep learning framework, and investigate several model architectures for better exploiting articulatory features.

## 2. ARTICULATORY FEATURES

### 2.1. Articulatory Feature Space

Articulatory features are the conventions of phonemes, which can reflect the pronunciation mechanisms and positions of articulators when pronouncing each phoneme, such as lip, tongue etc. Multiple streams are employed to describe articulatory features from different angles and each stream is abstracted to several classes to indicate different behaviors. According to relevant linguistic researches, we adopt the articulatory feature space proposed in [22], which is shown in Table 1.

### 2.2 Use of Articulatory Features for MDD

Since the pronunciation of each phoneme are decided by its production position and mechanism, articulatory features can help L2 learners correct their mispronunciation. Besides, thanks to the distinctive information included in articulatory features, the use of articulatory features in MDD is helpful in distinguishing phonemes, especially for some confusing phonemes pairs. Thus, in this paper, we employ the articulatory features to boost the state-of-the-art approaches to MDD and investigate several model architectures for better exploiting articulatory features in MDD.

**Table 2.** Expected British English articulatory features (part) [22]

| ***IPA*** | ***phoneme*** | ***jaw*** | ***lip separation*** | ***lip rounding*** | ***tongue frontness*** | ***tongue height*** | ***tongue tip*** | ***velum*** | ***voicing*** |
|---|---|---|---|---|---|---|---|---|---|
| ɑː | *aa* | 3 | 2 | 1 | 1 | 0 | 0 | 0 | 1 |
| æ | *ae* | 3 | 3 | 2 | 3 | 0 | 0 | 0 | 1 |
| ə | *ax* | 2 | 2 | 2 | 2 | 1 | 0 | 0 | 1 |
| aɪ | *ay* | 3 | 2 | 2 | 2 | 0 | 0 | 0 | 1 |
| | | 1 | 2 | 3 | 3 | 2 | 0 | 0 | 1 |
| ɛ | *eh* | 3 | 2 | 2 | 3 | 1 | 0 | 0 | 1 |
| ɪ | *ih* | 3 | 2 | 3 | 4 | 2 | 0 | 0 | 1 |
| u | *uw* | 1 | 1 | 0 | 1 | 3 | 0 | 0 | 1 |
| b | *b* | 1 | 0 | 2 | 2 | 1 | 1 | 0 | 1 |
| | | 1 | 2 | 2 | 2 | 1 | 1 | 0 | 1 |
| t | *t* | 1 | 1 | 2 | 4 | 3 | 4 | 0 | 0 |
| | | 1 | 2 | 2 | 4 | 2 | 3 | 0 | 0 |
| ð | *dh* | 2 | 2 | 2 | 4 | 2 | 2 | 0 | 1 |
| z | *z* | 1 | 2 | 2 | 3 | 3 | 3 | 0 | 1 |
| s | *s* | 1 | 2 | 2 | 3 | 3 | 3 | 0 | 0 |
| ʃ | *sh* | 2 | 2 | 1 | 3 | 3 | 0 | 0 | 0 |

### 2.3. Mapping Articulatory Features to Phone-State

The precise articulatory features are usually estimated by the electromagnetic articulography (EMA) and are expensive to obtain [23]. In our work, a mapping chart [22] is adopted to map phonemes to articulatory features. Part of the chart is shown in Table 2, and the full version can be found in [22]. Given dynamic process of speech, the articulatory features may change during the pronunciation of a phoneme, the mapping chart provides two sets of articulatory features for the start and end portions respectively of such phonemes, as illustrated by the /*ay*/,/*b*/ and /*t*/ lines in Table 2.

To utilize the articulatory features in MDD, they must be first mapped to corresponding speech frames. A 3-state GMM-HMM has been adopted to align the phoneme-level text transcriptions with the speech frames. For the speech frame of a particular phoneme, if it belongs to the first HMM state, the articulatory features at the start portion of the phoneme are used; if the frame is from the third state, the articulatory features at the end portion of the phoneme are used; otherwise the frame is from the second state, the interpolation values between the start and end portion articulatory features are used. An example of how to assign corresponding articulatory features is illustrated in Fig. 1.

## 3. ACOUSTIC-PHONEMIC MODEL WITH ARTICULATORY FEATURES

### 3.1. Acoustic-Phonemic Model (APM)

The acoustic-phonemic model (APM) is one of state-of-the-art approaches to MDD which calculates the phonetic posterior probabilities from the input acoustic and phonemic

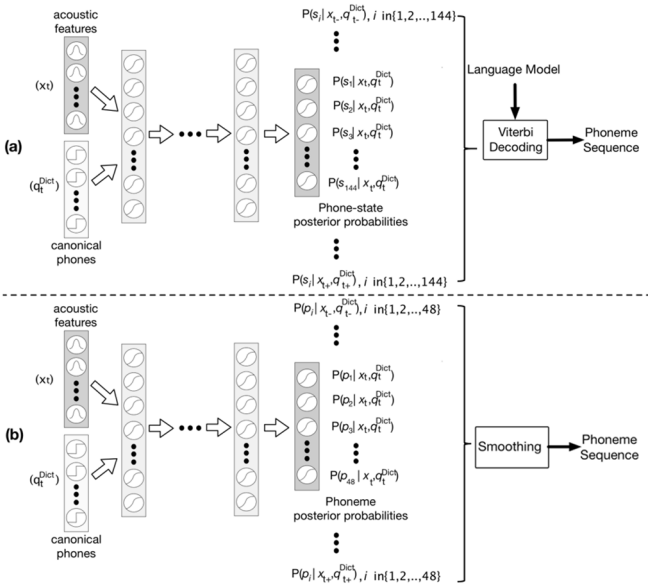


**Fig. 2**. (a) Diagram of state-based acoustic-phonemic model (APM); (b) Diagram of phoneme-based APM

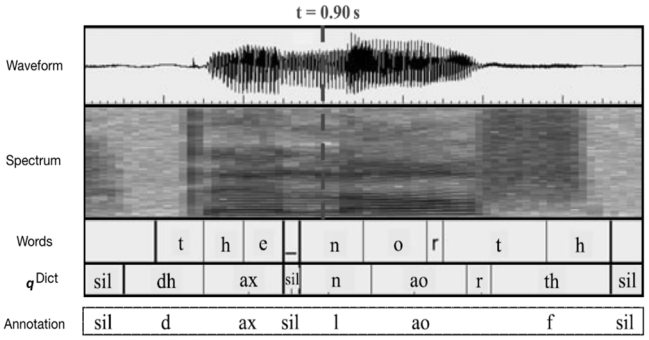


**Fig. 3**. An example of L2 speech aligned with canonical phones $q_t^{Dict}$ [13]

features [13]. Canonical phonemes are introduced as features in APM, adding context information to the input and deriving better performance over other methods. For each frame, APM uses Mel-frequency cepstral coefficients (MFCC) as acoustic features ($x_t$) and 7 canonical phonemes (3 before, 1 current and 3 after) as phonemic features ($q_t^{Dict}$), where the alignment between acoustic frame and the corresponding canonical phoneme is performed by GMM-HMM. For example in Fig. 3, the 7 canonical phonemes ($q_t^{Dict}$) of the frame *t*=0.90s are */dh ax sil n ao r th/*.

The APM proposed in [13] is state-based, which is shown in Fig. 2(a). The input of APM is the concatenation of $x_t$ and $q_t^{Dict}$ and after several hidden layers, the phone-state posterior probabilities $P(s_i|x_t,q_t^{Dict}), i \in [1 \dots 144]$ are derived. Finally, Viterbi decoding is used to generate the recognized phoneme sequence.

Considering the pronunciations in L2 speech are sometimes nonstandard and unordered, the pre-trained language model may limit the search space of the uncertain speech and influence the performance. So in this paper, we introduce a phoneme-based APM as baseline which is shown in Fig. 2(b). There are two differences: First, the phoneme-based APM is trained with phoneme labels that are transformed from state labels; second, the decoding is replaced with a smoothing process on the frame-level recognition results, during which a phoneme will be ignored if its duration is less than 2 frame-length.

### 3.2. Articulatory-Acoustic-Phonemic Model (AAPM)

Since the specific articulatory features can be used to identify phonemes, we incorporate them with APM and propose the articulatory-acoustic-phonemic model (APPM) that calculates the phoneme probabilities from input predicted articulatory features, acoustic features and phonemic features.

There are two stages in AAPM as shown in Fig. 4(a). In stage 1, an articulatory multi-task deep neural network (Articulatory-MT-DNN or A-MT-DNN) is used to predict feature in each articulatory stream, which calculates the class probabilities of 8 streams from input acoustic features ($x_t$) and phonemic features ($q_t^{Dict}$) with multi-task manner. The complete output probabilities of all tasks $P(A_k^j|x_t,q_t^{Dict}), k \in \{1,2,\dots,8\}, j \in \{1,\dots,|A_k|\}$ are treated as the predicted articulatory features input to stage 2. In the stage 2, the phoneme posterior probabilities are derived with the concatenation of $x_t$, $q_t^{Dict}$ and the predicted articulatory features.

Compared with APM, AAPM uses articulatory features as part of input, which are predicted through a well-trained A-MT-DNN and can provide the distinctive information between phonemes.

### 3.3. AAPM with Feature Representation (R-AAPM)

We introduce a feature representation module to AAPM, and adopt predicted articulatory features to re-represent original features instead of using them directly as input.

As shown in Fig. 4(b), the difference between R-AAPM and AAPM is the representation module. Same as AAPM, to estimate the class probabilities of each articulatory streams, an A-MT-DNN is trained first in stage 1. The trained A-MT-DNN is fixed during stage 2 training and the output probabilities of all streams $P(A_k^j|x_t,q_t^{Dict}), k \in \{1,2,\dots,8\}, j \in \{1,\dots,|A_k|\}$ are computed through the fixed A-MT-DNN. Then a dense output vector is derived through linear transformation in the dense layer which has equal length with the input features. Finally, the represented new features are obtained by adding input features and corresponding bits in the dense output vector.

Take *voicing* (i.e. $A_8$) as an example to demonstrate the effect of representation module. Let $f^i$ be the $i$th bit of input feature, and after the representation module, $f^i$ turns into $new\_f^i$, where $P(A_8^1|x,q^{Dict}), P(A_8^2|x,q^{Dict})$ is the probability of voiced and unvoiced; $w_{iv}, w_{iu}$ are corresponding parameters in dense layer:

$$new\_f^i = f^i + w_{iv}P(A_8^1|x,q^{Dict}) + w_{iu}P(A_8^2|x,q^{Dict}) \quad (1)$$

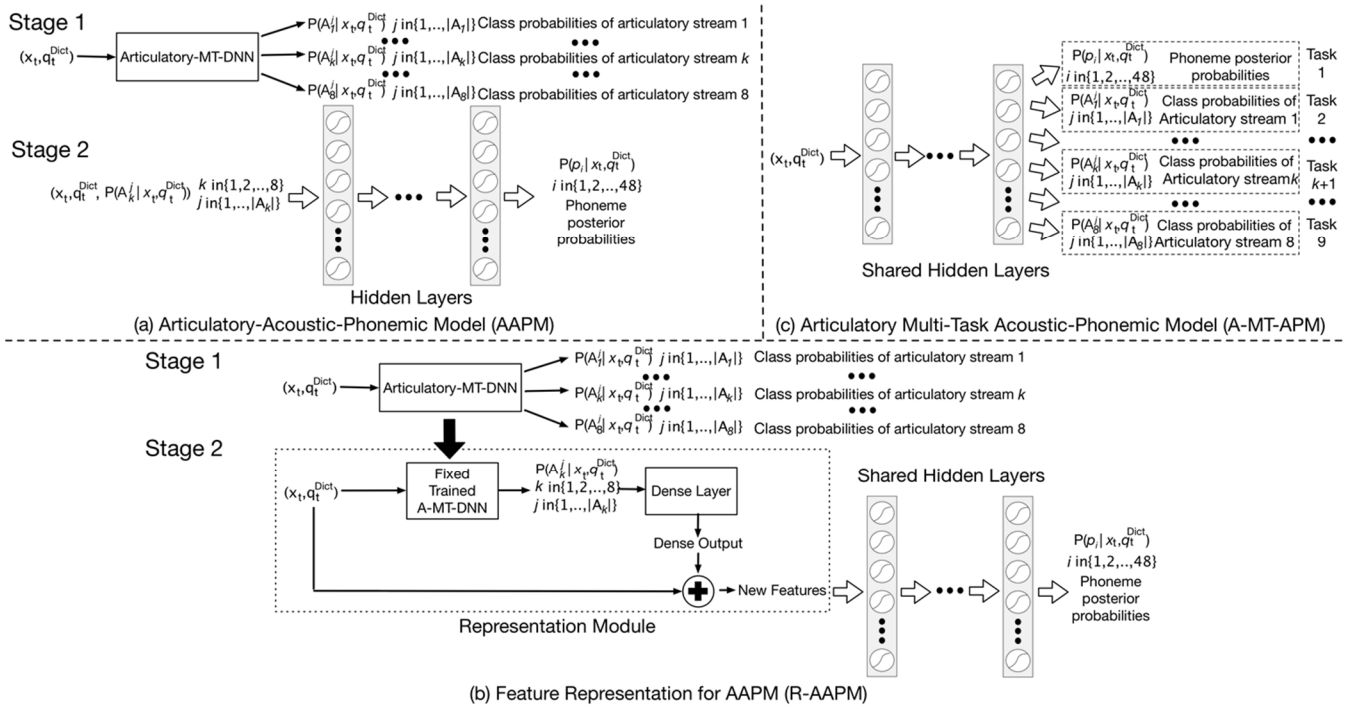


**Fig. 4.** Diagrams of the proposed (a) AAPM, (b) R-AAPM and (c) A-MT-APM

$$P(A_8^1|x, q^{Dict}) + P(A_8^2|x, q^{Dict}) = 1 \quad (2)$$

The mathematical expectation and variance after representation can be derived as:

$$\begin{aligned} E(new\_f^i) &= E\left(f^i + w_{iv}P(A_8^1|x, q^{Dict}) + w_{iu}P(A_8^2|x, q^{Dict})\right) \\ &= E(f^i) + w_{iv}E\left(P(A_8^1|x, q^{Dict})\right) + w_{iu}E\left(P(A_8^2|x, q^{Dict})\right) \\ &= E(f^i) + (w_{iv} - w_{iu})E(P(A_8^1|x, q^{Dict})) + w_{iu} \end{aligned} \quad (3)$$

$$\begin{aligned} D(new\_f^i) &= D\left(f^i + w_{iv}P(A_8^1|x, q^{Dict}) + w_{iu}P(A_8^2|x, q^{Dict})\right) \\ &= D\left(f^i + (w_{iv} - w_{iu})P(A_8^1|x, q^{Dict}) + w_{iu}\right) \\ &= D(f^i) + w_i D\left(P(A_8^1|x, q^{Dict})\right) + 2w_i cov\left(f^i, P(A_8^1|x, q^{Dict})\right) \end{aligned} \quad (4)$$

Suppose $w_i = w_{iv} - w_{iu}$ and $P(A_8^1|x, q^{Dict})|voi \sim N(\mu_v, \sigma_v^2)$, $P(A_8^1|x, q^{Dict})|unv \sim N(\mu_u, \sigma_u^2)$:

$$E(new\_f^i|voi) = E(f^i|voi) + w_i\mu_v + w_{iu} \quad (5)$$

$$E(new\_f^i|unv) = E(f^i|unv) + w_i\mu_u + w_{iu} \quad (6)$$

$$\begin{aligned} D(new\_f^i|voi) &= D(f^i|voi) + w_i\sigma_v^2 + 2w_i\left(E\left(f^i P\left(A_8^1|x, q^{Dict}\right)\right) - \mu_v E(f^i)\right) \\ &= D(f^i|voi) + w_i\sigma_v^2 + 2w_i E\left(f^i(P\left(A_8^1|x, q^{Dict}\right) - \mu_v)\right) \end{aligned} \quad (7)$$

$$D(new\_f^i|unv) = D(f^i|unv) + w_i\sigma_u^2 + 2w_i E\left(f^i(P\left(A_8^1|x, q^{Dict}\right) - \mu_u)\right) \quad (8)$$

If reliable $P(A_8^1|x_t, q_t^{Dict}), P(A_8^2|x_t, q_t^{Dict})$ can be obtained, $\sigma_v^2$, $\sigma_u^2$ are near to 0 while $\mu_v$, $\mu_u$ are near to 1 and 0 respectively. So:

$$E(new\_f^i|voi) \to E(f^i|voi) + w_{iv} \quad (9a)$$

$$E(new\_f^i|unv) \to E(f^i|unv) + w_{iu} \quad (9b)$$

$$D(new\_f^i|voi) \to D(f^i|voi) \quad (9c)$$

$$D(new\_f^i|unv) \to D(f^i|unv) \quad (9d)$$

Hence, the representation module will change the feature distribution distance between different classes among a specific articulatory stream and almost keep the feature distribution within intra-class, which may use articulatory feature information more effectively and further enhance the model performance.

**Table 3**. The definitions in hierarchical evaluation

| For all Phonetic Unit | | Recognition Result: Correct Pronunciation | Recognition Result: Mispronunciation |
|---|---|---|---|
| Manually Transcribed Phonetic Unit | Correct Pronunciation | **TA** | **FR** |
| | Mispronunciation | **FA** | **TR (CD/DE)** |

### 3.4 Articulatory Multi-Task Acoustic-Phonemic Model (A-MT-APM)

Apart from adding predicted articulatory features into input features, in order to take advantage of articulatory features in the training stage directly, we also propose an articulatory multi-task acoustic-phonemic model (A-MT-APM) by incorporating the multi-task learning technique into APM, where the phoneme recognizer and classifiers for articulatory streams are trained together.

As shown in Fig. 4(c), 9 tasks are involved in A-MT-APM: Task 1 is a free phoneme recognizer and the remaining 8 tasks are classifiers for all articulatory streams. These tasks share equal importance and are trained simultaneously in multi-task learning manner. The acoustic and phonemic features are concatenated as input, and after several shared hidden layers jointly trained by all tasks, the 9 separate layers computing corresponding results are the output of A-MT-APM.

Different from AAPM and R-AAPM, A-MT-APM treats articulatory feature predictions as subtasks instead of using them as inputs. It can boost the performance by jointly

**Table 4**. Experimental results of phoneme recognition and MDD with different metrics

| Method | *Performance of Recognition* | | *Performance of Mispronunciation Detection and Diagnosis* | | | | |
|---|---|---|---|---|---|---|---|
| | *Correct* | *Accuracy* | *Precision* | *Recall* | *F1-measure* | *Detection Accuracy* | *Diagnostic Accuracy* |
| State-based APM | 81.4% | 76.3% | 63.4% | **83.7%** | 72.1% | 89.0% | 68.4% |
| Phoneme-based APM | 89.8% | 83.4% | 71.8% | 80.2% | 75.7% | 92.1% | 77.3% |
| AAPM | 89.0% | 82.9% | 72.9% | 80.9% | 76.7% | 92.5% | 77.4% |
| R-AAPM | 90.6% | 85.2% | 74.2% | 79.3% | 76.7% | 92.7% | 79.5% |
| A-MT-APM | **93.3%** | **87.2%** | **86.5%** | 76.7% | **81.3%** | **94.6%** | **84.3%** |

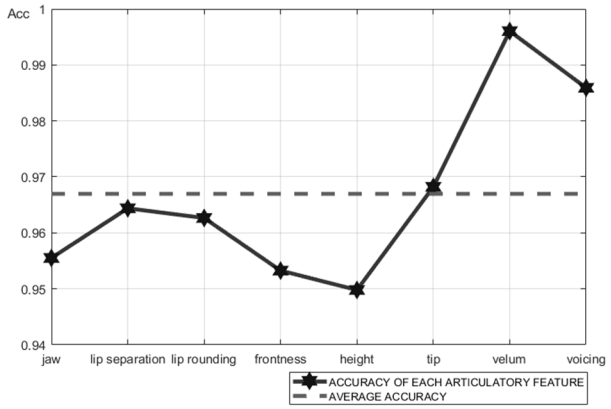


**Fig. 5.** The accuracy of articulatory feature prediction

training and avoiding the impact of deviations from previous articulatory feature predictions.

## 4. EXPERIMENTS

### 4.1. Speech Corpus

CU-CHLOE-C (**C**hinese **U**niversity **Ch**inese **L**earners of **E**nglish - **C**antonese speakers) is used in our experiments, which contains L2 English speech uttered by 100 Cantonese speakers. 30% audios of each speaker are labeled by skilled linguists with their actual pronunciations (i.e. Annotation in Fig. 3).

### 4.2. Experimental Setup

Labelled audios from 80 speakers in CHLOE-C are picked as training set (9.5 hours) and the labelled audios from remaining 20 speakers are selected as test set (2.0 hours).

11 frames (5 before, 1 current and 5 after) of MFCC are used as the acoustic features $x_t$. 7 canonical phonemes (3 before, 1 current and 3after) are applied as the phonemic features ($q_t^{Dict}$). 11 frames (5 before, 1 current and 5 after) of complete outputs derived through the trained A-MT-DNN are employed as predicted articulatory features.

Five different models are involved in our experiments: (1) Baseline state-based APM, (2) Baseline phoneme-based APM (3) AAPM, (4) R-AAPM and (5) A-MT-APM: (1) and (2) are implemented first to select a better baseline. Then (3), (4) and (5) are built to verify whether articulatory features will improve model performance or not and how to exploit articulatory features more effectively.

**Table 5**. Confusion matrices of frequently misrecognized phonemes

| | | Annotation (Vowels) | | | | | | | Annotation (Consonants) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ae | ah | ax | ay | ih | ix | | d | dh | t | s | sh | z |
| Baseline APM | ae | 923 | 1 | 8 | 2 | 0 | 0 | d | 699 | 46 | 9 | 3 | 0 | 0 |
| | ah | 7 | 415 | 6 | 1 | 1 | 1 | dh | 14 | 462 | 0 | 1 | 0 | 1 |
| | ax | 51 | 6 | 1102 | 31 | 8 | 14 | t | 63 | 2 | 3025 | 3 | 2 | 1 |
| | ay | 1 | 0 | 2 | 807 | 2 | 0 | s | 0 | 0 | 13 | 1680 | 4 | 21 |
| | ih | 0 | 0 | 1 | 1 | 658 | 21 | sh | 0 | 0 | 16 | 6 | 359 | 0 |
| | ix | 0 | 3 | 18 | 9 | 90 | 434 | z | 0 | 6 | 1 | 21 | 0 | 365 |
| A-MT-APM | ae | 926 | 2 | 3 | 1 | 1 | 0 | d | 803 | 49 | 9 | 0 | 0 | 0 |
| | ah | 5 | 427 | 5 | 1 | 1 | 1 | dh | 7 | 532 | 0 | 0 | 0 | 0 |
| | ax | 40 | 10 | 1254 | 21 | 7 | 13 | t | 41 | 1 | 3140 | 6 | 0 | 1 |
| | ay | 0 | 0 | 0 | 819 | 0 | 1 | s | 1 | 0 | 10 | 1701 | 2 | 15 |
| | ih | 1 | 0 | 2 | 0 | 695 | 20 | sh | 1 | 0 | 10 | 4 | 356 | 0 |
| | ix | 1 | 2 | 13 | 4 | 59 | 478 | z | 0 | 6 | 1 | 17 | 0 | 365 |

After preliminary experiments, all the models above contain seven hidden layers with 2048 units per layer and *tanh* as activation function. The A-MT-DNN in (3), (4) are determined to have five hidden layers with 1024 units per layers and *tanh* as activation function.

### 4.3. Performance of Articulatory Feature Prediction

In AAPM and R-AAPM, articulatory features are predicted through A-MT-DNN first. The articulatory feature prediction accuracy of each articulatory stream is shown in Fig. 5. All accuracies are above 0.950 and the average accuracy is 0.967, so it can be assumed that reliable articulatory feature predictions can be obtained with A-MT-DNN.

### 4.4. Performance of MDD

The performance of free phoneme recognition is evaluated with the correctness and accuracy which are computed against linguist's annotations:

$$Correct = \frac{N-S-D}{N}, Accuracy = \frac{N-S-D-I}{N} \quad (10)$$

where $N$ is the total number of phonemes, and $S, D, I$ are the counts of substitutions, deletions and insertions.

The performance of MDD is assessed with the hierarchical evaluation structure proposed in [24]. True Acceptance (TA), True Rejection (TR), False Rejection (FR), False Acceptance (FA), Correct Diagnosis (CD) and Diagnosis Error (DE) are defined as Table 3. Precision, Recall, F1-Measure and accuracies of mispronunciation detection and diagnosis which are computed as follows:

$$Precision = \frac{TR}{TR + FR}, Recall = \frac{TR}{TR + FA} \quad (11a)$$

$$F1 - measure = 2\frac{Precision \times Recall}{Precision + Recall} \quad (11b)$$

$$DetectionAccuracy = \frac{TA + TR}{TA + FR + FA + TR} \quad (11c)$$

$$DiagnosticAccuracy = \frac{CD}{CD + DE} \quad (11d)$$

The results are listed in Table 4. Compared with state-based APM, phoneme-based APM achieves a great improvement. So phoneme-based APM is set as baseline and all other experiments are conducted with phoneme-based output layer. According to experimental results, we can find the AAPM, R-AAPM and A-MT-APM show obvious improvements. Especially with A-MT-APM, the F1-Measure and diagnostic accuracy are greatly improved to 81.3% from 75.7% and 84.3% from 77.3% respectively.

Besides, confusion matrices of frequently misrecognized phonemes by phoneme-based APM are shown in Table 5. We can find that the recognition between confusable pairs is greatly improved by our proposed A-MT-APM.

In summary, a better MDD system can be achieved with the assistance of articulatory features.

## 5. CONCLUSION

This paper proposes novel approaches to MDD by introducing articulatory features to APM. Three models with articulatory features are proposed: (1) AAPM; (2) R-AAPM; (3) A-MT-APM. Compared with the baseline phoneme-based APM, our proposed approaches perform better in all metrics. Specifically, the A-MT-APM approach gains 5.6% and 7.0% improvement in F1-Measure and diagnostic accuracy respectively. In conclusion, APM with articulatory features improves the MDD performance.

## 6. REFERENCES


[1] Jo, C. H., Kawahara, T., Doshita, S., and Dantsuji, M., "Automatic pronunciation error detection and guidance for foreign language learning.", *Fifth International Conference on Spoken Language Processing*, 1998.

[2] Franco, H., Neumeyer, L., Ramos, M., and Bratt, H., "Automatic detection of phone-level mispronunciation for language learning.", *Sixth European Conference on Speech Communication and Technology*, 1999.

[3] Witt, S. M., and Young, S. J., "Phone-level pronunciation scoring and assessment for interactive language learning.", *Speech communication* 30.2 (2000), pp.95-108, 2000

[4] Menzel, W., Herron, D., Bonaventura, P., and Morton, R., "Automatic detection and correction of non-native English pronunciations.", *Proceedings of INSTILL*(2000): pp.49-56, 2000

[5] Seneff, S., Wang, C., and Zhang, J., "Spoken conversational interaction for language learning.", *InSTIL/ICALL Symposium* 2004, 2004.

[6] Zheng, J., Huang, C., Chu, M., Soong, F. K., and Ye, W. P., "Generalized segment posterior probability for automatic mandarin pronunciation evaluation.", *ICASSP 2007*. IEEE International Conference on. Vol. 4, pp. IV-201, 2007.

[7] Truong, K., Neri, A., Cucchiarini, C., and Strik, H., "Automatic pronunciation error detection: an acoustic-phonetic approach.", *InSTIL/ICALL Symposium 2004*, 2004.

[8] Strik, H., Truong, K., De Wet, F., and Cucchiarini, C., "Comparing different approaches for automatic pronunciation error detection.", *Speech communication*51.10 (2009), pp.845-852, 2009.

[9] Lee, A., and Glass, J. R., "Context-dependent pronunciation error pattern discovery with limited annotations.", *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[10] Harrison, A. M., Lo, W. K., Qian, X., and Meng, H., "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training.", *SLaTE*, 2009.

[11] Lo, W. K., Zhang, S., and Meng, H., "Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system.", *11th Annual Conference of the International Speech Communication Association*, 2010.

[12] Qian, X., Soong, F. K., and Meng, H., "Discriminative acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (CAPT).", *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[13] Li, K., Qian, X., and Meng, H. "Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks.", *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.1 (2017), pp.193-207, 2017.

[14] Yu, D., and Deng, L., "Automatic speech recognition: A deep learning approach.", *Springer*, 2014.

[15] LeCun, Y., Bengio, Y., and Hinton, G., "Deep learning.", *Nature* 521.7553 (2015), pp.436-444, 2015.

[16] Caruana, R., "Multitask learning.", Learning to learn. *Springer* US, pp.95-133, 1998.

[17] P. Ladefoged*, A Course in Phonetics*, 5th ed. Boston, MA: Thomson Wadsworth, 2006.

[18] Li, W., Li, K., Siniscalchi, S. M., Chen, N. F., and Lee, C. H. "Detecting Mispronunciations of L2 Learners and Providing Corrective Feedback Using Knowledge-Guided and Data-Driven Decision Trees. " *Interspeech*. 2016, pp. 3127-3131, 2016.

[19] J. Sun and L. Deng, "An overlapping-feature-based phonological model incorporating linguistic constraints: Applications to speech recognition," *J. Acoust. Soc. Amer.*, vol. 111, no. 2, pp. 1086–1101, 2002.

[20] M. Richardson, J. Bilmes, and C. Diorio, "Hidden-articulator Markov models for speech recognition, " *Speech communication*., vol. 41, no. 2, pp.511–529, Oct. 2003.

[21] J. Tepperman and S. Narayanan, "Hidden-articulator Markov models for pronunciation evaluation, " in *Proc. ASRU*, San Juan, Puerto Rico, 2005, pp. 174–179.

[22] J. Tepperman and S. Narayanan. "Using articulatory representations to detect segmental errors in nonnative pronunciation." *IEEE Transactions on audio, speech, and language processing* 16.1 (2008), pp.8-22, 2008.

[23] Sun L, Li K, Wang H, et al. "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training", ICME, 2016 IEEE International Conference on. IEEE, pp. 1-6, 2016.

[24] Qian, X., Meng, H., and Soong, F., "Capturing l2 segmental mispronunciations with joint-sequence models in computer-aided pronunciation training (CAPT).", ISCSLP, 2010 7th International Symposium on IEEE, pp. 84-88, 2010.