APPLYING MULTITASK LEARNING TO ACOUSTIC-PHONEMIC MODEL FOR MISPRONUNCIATION DETECTION AND DIAGNOSIS IN L2 ENGLISH SPEECH

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

[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, hmmeng}@se.cuhk.edu.hk, clh-dcs@tsinghua.edu.cn

ABSTRACT

For mispronunciation detection and diagnosis (MDD), nowadays approaches generally treat the phonemes in correct and mispronunciations as the same despite the fact they may actually carry different characteristics. Furthermore, serious data imbalance issue between correct and mispronunciation in dataset further influences the performances. To address these problems, this paper investigates the use of multi-task (MT) learning technique to enhance the acoustic-phonemic model (APM) for MDD. The phonemes in correct and mispronunciations are processed separately but in multi-task manner considering both correct and mispronunciation recognition tasks. A feature representation module is further proposed to improve performance. Compared with baseline APM, the proposed MT-APM, R-MT-APM achieve better performance not only in Precision, Recall and F-Measure, but also in mispronunciation detection and diagnosis accuracies. With feature representation module, R-MT-APM achieves the highest mispronunciation detection accuracy.

Index Terms— Computer-aided pronunciation training, mispronunciation detection and diagnosis, multi-task learning, acoustic-phonemic model, feature representation

1. INTRODUCTION

By providing self-learning opportunities in second-language (L2) learning, computer-aided pronunciation training (CAPT) systems have attracted wide research interests. As the core of CAPT, mispronunciation detection and diagnosis (MDD) aims at detecting mispronunciations in the learner's speech and providing diagnosis feedback.

Several methods have been proposed for MDD [1]-[15] which can be grouped into two categories. The first is pronunciation scoring based approaches that use different types of confidence measures, such as likelihood ratios, phone posterior probabilities, etc. to evaluate the scores of pronunciations and give scores as feedback [2]-[7]. Goodness of pronunciation (GOP) is the most representative score measure. GOP is obtained from the target phoneme's posterior probability computed by the acoustic models [3]. Scoring methods can make a judgement on whether the pronunciation is correct or wrong by thresholding, but cannot give diagnostic feedback. The second category of methods aims at detecting the details of mispronunciations and giving feedback about specific errors such as phoneme substitutions, deletions and insertions [8]-[12]. Extended recognition networks (ERN) perform well in diagnosing mispronunciation types [13]-[15], which incorporate manually designed or data-derived phonological rules to generate possible phonemic paths in a word, including the canonical phonemic path and common mispronunciation paths. However, such approaches lack the ability to diagnose mispronunciation patterns in L2 speech that are not covered by the phonological rules. The acoustic-phonemic model (APM) is another approach to MDD, which is a deep neural network (DNN) that maps input features with acoustic information and phonemic context information into phonemic posterior probabilities [12]. APM can achieve better performance in MDD due to the incorporation of the canonical phonemic context information.

Riding on the development of automatic speech recognition (ASR) [16] and deep learning technologies [17], MDD has achieved much improvement, but there are still problems in current approaches. First, the phonemes in mispronunciation may carry characteristics different from the counterparts in correct pronunciation. For example, a common mispronunciation found in Cantonese L2 English shows that the phoneme /th/ may be mispronounced as a sound that bears resemblance to both /f/ and /th/. But the current systems only regard them as the same. Second, the proportion of mispronunciation is far less than correct in L2 speech [12]. Such data imbalance problem will influence the performance of the systems.

Inspired by multi-task learning in ASR [18]-[23], we propose the multi-task APM to solve the above problems. Correct and mispronunciations are processed separately in multi-task manner.

2. ACOUSTIC-PHONEMIC MODEL (APM)

The acoustic-phonemic model (APM) calculates the phone-state posterior probabilities from the input acoustic and phonemic features [12]. As shown in Fig. 1, the input of APM is the concatenation of acoustic features (x_t) and phonemic features (q_t^{Dict}). After several hidden layers, the phone-state posterior probabilities $\mathrm{P}(s_i|x_t, q_t^{Dict}), i \in [1 \dots 144]$ are derived. Finally, Viterbi decoding is used to generate the recognized phoneme sequence. For each frame, APM uses Mel-frequency cepstral coefficients (MFCC) as acoustic features (x_t) and 7 canonical phones (3 before, 1 current and

3 after) as phonemic features (q_t^{Dict}), where the alignment between acoustic frame and the corresponding canonical phone is performed by Gaussian mixture model-hidden Markov model (GMM-HMM). For example, in Fig.2(a), the 7 canonical phones (q_t^{Dict}) of the frame *t*=0.90s are */dh ax sil n ao r th/*.

APM introduces canonical phones as features and adds context information to input, leading to better results over other methods. However, APM tends to miss the recall of actually mispronounced phones (recall rate is less than 70%) [12]. One reason is that the proportion of mispronunciations is far less than the correct pronunciations (around 16:84 according to dataset statistics). Such data imbalance problem degrades the performance of the model. Furthermore, there may be differences between correct and mispronunciations of a phone. Treating them as the same may further affect model's performance.

3. MULTI-TASK LEARNING FOR APM

3.1. Multi-Task APM (MT-APM)

To solve the aforementioned problems, we propose the multi-task APM (MT-APM) by incorporating the multi-task learning technique into APM, where the phone state posterior probabilities are trained by multi-task learning considering both correct and mispronunciation recognition tasks.

3.1.1. Structure of MT-APM

As shown in Fig.3(a), MT-APM involves two tasks: Task 1 deals with correct pronunciations (correct recognizer) and Task 2 for mispronunciations (mispronunciation recognizer). The two tasks are of equal importance and can be trained and used with multi-task learning. The acoustic and phonemic features (x_t, q_t^{Dict}) are concatenated together as the input of MT-APM; followed by several shared hidden layers that are jointly trained by the two tasks; and the output of MT-APM are two separate layers computing phone-state posterior probabilities for the two tasks respectively. In MT-APM, the correct and mispronunciations are processed separately in two tasks but with shared common characteristics in multi-task learning, which improves the recognition performance (in calculating phone-state posterior probabilities) of both correct and mispronunciation recognizers.

3.1.2. Labels of Data for MT-APM

For APM, there are 144 phoneme states (48 phonemes * 3 states per phoneme) for labels. The label of each frame is obtained by forced alignment with linguists' *Annotation*. In MT-APM, we introduce two new states (*mis* and *cor*) for the two tasks, *mis* for Task 1 and *cor* for Task 2 respectively. *mis* serves as the "collection" state for all pronunciations that do not belong to any of the 144 states of correct pronunciations in correct recognizer (Task 1), while *cor* serves as similar purpose in mispronunciation recognizer (Task 2). As shown in Fig.2, the *Annotations* are compared with the canonical phones q_t^{Dict} at frame level to determine the correct and mispronunciation segments. If a frame belongs to a correct segment (i.e. *Annotation* is the same as canonical phone), its label for Task 1 is the canonical phone, while its label for Task 2 is *cor*. If a frame belongs to a mispronounced segment (i.e. *Annotation* is different from canonical phone), its label for Task 1 is *mis*, while its label for Task 2 is the *Annotation* phone. Fig.2(b) provides the example labels of the data for MT-APM.

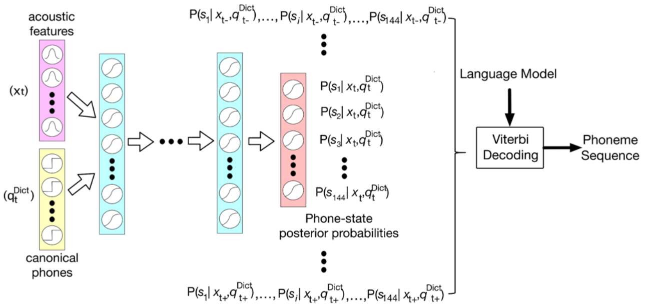

acoustic features
(xt)
(q_t^Dict)
canonical phones
P(s1| xt-,q_t-^Dict),…, P(si| xt-,q_t-^Dict),…,P(s144| xt-,q_t-^Dict)
P(s1| xt,qt^Dict)
P(s2| xt,qt^Dict)
P(s3| xt,qt^Dict)
P(s144| xt,qt^Dict)
Phone-state posterior probabilities
P(s1| xt+,q_t+^Dict),…, P(si| xt+,q_t+^Dict),…,P(s144| xt+,q_t+^Dict)
Language Model
Viterbi Decoding
Phoneme Sequence

Fig.1. Diagram of acoustic-phonemic model (APM)

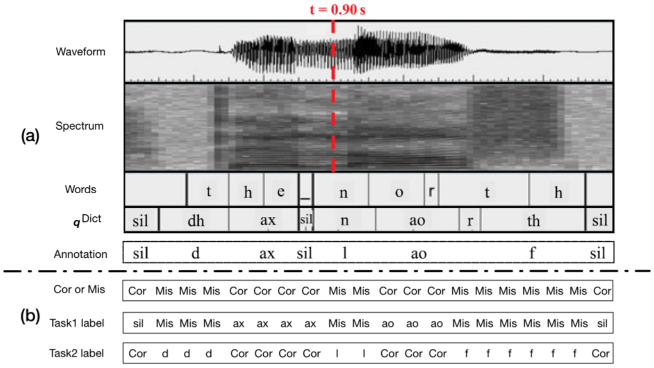


Fig.2. (a) An example of L2 speech aligned with canonical phones q_t^{Dict} [12]; and (b) corresponding ground-truth labels in MT-APM and R-MT-APM

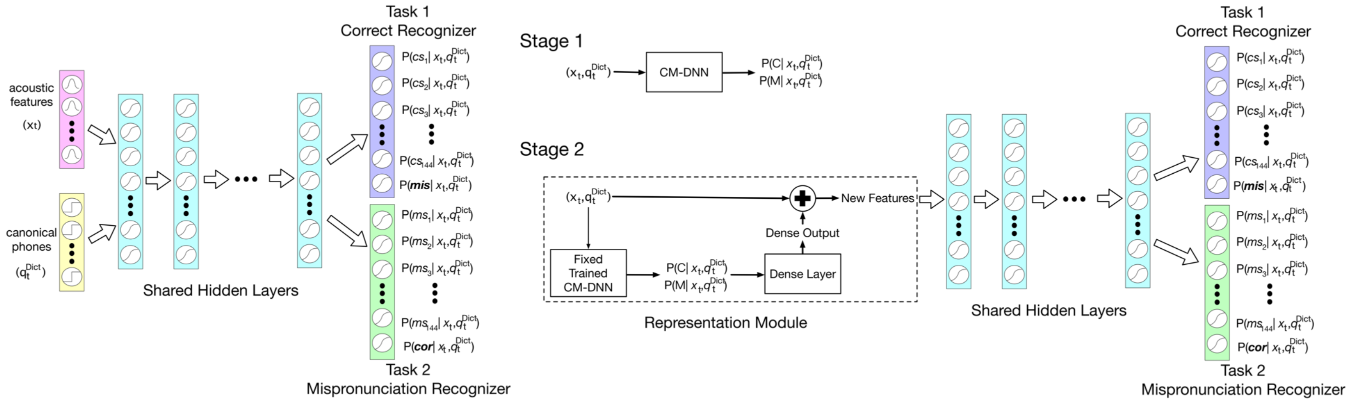
3.1.3. Joint Decoding for MT-APM

For Task 1, its output is a vector representing the probabilities of all correctly pronounced phone-states $\mathrm{P}(cs_i|x_t,q_t^{Dict})$ and the probability of mispronunciation $\mathrm{P}(mis|x_t,q_t^{Dict})$. While for Task 2, its output is the probabilities of all mispronounced phone-states $\mathrm{P}(ms_i|x_t,q_t^{Dict})$ and the probability of correct pronunciation $\mathrm{P}(cor|x_t,q_t^{Dict})$. Here, cs_i is the ith correctly pronounced phone-state and ms_i is the ith mispronounced phone-state, $i \in [1 \dots 144]$.

Joint decoding scheme is then proposed for MT-APM, as shown in Fig. 4. For each frame, the $\mathrm{P}(mis|x_t,q_t^{Dict})$ from Task 1 and $\mathrm{P}(cor|x_t,q_t^{Dict})$ from Task 2 are compared first. If $\mathrm{P}(mis|x_t,q_t^{Dict})$ is greater than $\mathrm{P}(cor|x_t,q_t^{Dict})$, the frame is treated as a mispronounced frame and the $\mathrm{P}(ms_i|x_t,q_t^{Dict}), i \in [1 \dots 144]$ from Task 2 are used as the final 144 bits output for Viterbi decoding. Otherwise, the frame is treated as correctly pronounced and the $\mathrm{P}(cs_i|x_t,q_t^{Dict}), i \in [1 \dots 144]$ are used for decoding.

3.2. Feature Representation for MT-APM (R-MT-APM)

In MT-APM, the two tasks recognize the frames into different phone-states (cs_i or ms_i) according to the distribution of the input acoustic and phonemic features (x_t, q_t^{Dict}). To increase



(a) MT-APM

(b) R-MT-APM

Fig.3. Diagrams of the proposed MT-APM (a) and R-MT-APM (b)

the distinction between feature distributions of correct and mispronunciations, we improve MT-APM by adding the feature representation module and propose the R-MT-APM.

3.2.1. Structure of R-MT-APM

As shown in Fig.3(b), the difference between R-MT-APM and MT-APM is the representation module. To derive R-MT-APM, a correct-mispronunciation DNN (CM-DNN) is first trained in stage 1, to judge if the current frame has correct pronunciation (*cor*) or mispronunciation (*mis*). The trained CM-DNN is fixed during the network training of stage 2. For input features (x_t, q_t^{Dict}), $\mathrm{P}(C|x_t,q_t^{Dict})$ and $\mathrm{P}(M|x_t,q_t^{Dict})$ are derived first through the fixed trained CM-DNN. A dense output vector is then computed which has the same bits as input features through linear transformation in the dense layer. The represented new features are finally computed by adding corresponding bits of input features (x_t, q_t^{Dict}) and the dense output vector.

3.2.2. Proof of the Effectiveness of Representation Module

In R-MT-APM, the representation module uses information about whether current frame has correct pronunciation or mispronunciation, as computed by a well-trained CM-DNN. Such a process is expected to increase the inter-class distance while keeping the intra-class characteristics of feature distributions of correct pronunciations and mispronunciations, and hence will further boost the performance of the model.

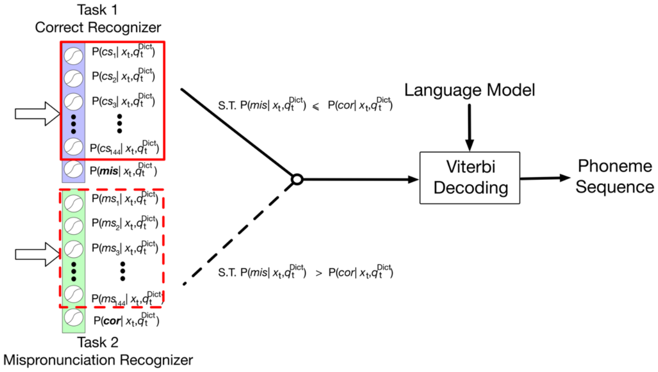

Task 1
Correct Recognizer
$P(cs_1|x_t,q_t^{Dict})$
$P(cs_2|x_t,q_t^{Dict})$
$P(cs_3|x_t,q_t^{Dict})$
$P(cs_{144}|x_t,q_t^{Dict})$
$P(\textbf{mis}|x_t,q_t^{Dict})$
$P(ms_1|x_t,q_t^{Dict})$
$P(ms_2|x_t,q_t^{Dict})$
$P(ms_3|x_t,q_t^{Dict})$
$P(ms_{144}|x_t,q_t^{Dict})$
$P(\textbf{cor}|x_t,q_t^{Dict})$
Task 2
Mispronunciation Recognizer
S.T. $P(mis|x_t,q_t^{Dict}) < P(cor|x_t,q_t^{Dict})$
S.T. $P(mis|x_t,q_t^{Dict}) > P(cor|x_t,q_t^{Dict})$
Language Model
Viterbi Decoding
Phoneme Sequence

Fig.4. Joint decoding scheme for MT-APM

This section proofs that the proposed representation module satisfies such requirements.

Let f_t^i be the ith bit of input feature at time step t. After the representation module, f_t^i becomes $new\_f_t^i$:

$$new\_f_t^i = f_t^i + w_{i1}\mathrm{P}\left(C|x_t,q_t^{Dict}\right) + w_{i2}\mathrm{P}\left(M|x_t,q_t^{Dict}\right) \qquad (1)$$

where w_{i1} and w_{i2} are the parameters in the dense layer, and

$$\mathrm{P}\left(C|x_t,q_t^{Dict}\right) + \mathrm{P}\left(M|x_t,q_t^{Dict}\right) = 1 \qquad (2)$$

The mathematical expectation and variance of $new\_f_t^i$ can be computed as:

$$\begin{aligned} E(new\_f^i) &= E\left(f^i + w_{i1}P(C|x,q^{Dict}) + w_{i2}P(M|x,q^{Dict})\right) \\ &= E(f^i) + w_{i1}E\left(P(C|x,q^{Dict})\right) + w_{i2}E\left(P(M|x,q^{Dict})\right) \\ &= E(f^i) + (w_{i1} - w_{i2})E\left(P(C|x,q^{Dict})\right) + w_{i2} \qquad (3) \end{aligned}$$

$$\begin{aligned} D(new\_f^i) &= D\left(f^i + w_{i1}P(C|x,q^{Dict}) + w_{i2}P(M|x,q^{Dict})\right) \\ &= D(f^i + (w_{i1} - w_{i2})P(C|x,q^{Dict}) + w_{i2}) \\ &= D(f^i) + w_iD\left(P(C|x,q^{Dict})\right) + 2w_icov\left(f^i, P(C|x,q^{Dict})\right) \end{aligned} \qquad (4)$$

Assume $w_{i1} - w_{i2} = w_i$ and $P(C|x,q^{Dict})|Cor \sim N(\mu_C, \sigma_C{}^2)$, $P(C|x,q^{Dict})|Mis \sim N(\mu_M, \sigma_M{}^2)$:

$$E(new\_f^i|Cor) = E(f^i|Cor) + w_i\mu_C + w_{i2} \qquad (5)$$

$$E(new\_f^i|Mis) = E(f^i|Mis) + w_i\mu_M + w_{i2} \qquad (6)$$

$$\begin{aligned} D(new\_f^i|Cor) &= D(f^i|Cor) + w_i\sigma_C{}^2 + 2w_i(E\left(f^iP(C|x,q^{Dict})\right) - \mu_CE(f^i)) \\ &= D\left(f^i|Cor\right) + w_i\sigma_C{}^2 + 2w_iE\left(f^i(P\left(C|x,q^{Dict}\right) - \mu_C)\right) \qquad (7) \end{aligned}$$

$$D(new\_f^i|Mis) = D(f^i|Mis) + w_i\sigma_M{}^2 + 2w_iE(f^i(P(C|x,q^{Dict}) - \mu_M)) \qquad (8)$$

For CM-DNN, it is a binary classification (*cor* or *mis*) network. By careful and sufficient training, the variances of its outputs $\mathrm{P}(C|x_t,q_t^{Dict})$ and $\mathrm{P}\left(M|x_t,q_t^{Dict}\right)$ (i.e. $\sigma_C{}^2$, $\sigma_M{}^2$) can be controlled, while the expectations (i.e. μ_C, μ_M) are near to 1 and 0 respectively. Hence:

$$E(new\_f^i|Cor) \rightarrow E(f^i|Cor) + w_{i1} \qquad (9a)$$

$$E(new\_f^i|Mis) \rightarrow E(f^i|Mis) + w_{i2} \qquad (9b)$$

$$D\left(new\_f^i|Cor\right) \rightarrow D\left(f^i|Cor\right) \qquad (9c)$$

$$D\left(new\_f^i|Mis\right) \rightarrow D\left(f^i|Mis\right) \qquad (9d)$$

which indicates the feature representation module satisfies the above requirements.

4. EXPERIMENTS

4.1. Speech Corpus

Our experiments are based on the CU-CHLOE (**C**hinese **U**niversity **CH**inese **L**earners **o**f **E**nglish) corpus that contains L2 English speech uttered by 100 Cantonese

Table 1. Experimental results of phone recognition and MDD with different metrics

Dataset	Method	*Performance of Recognition*		*Performance of Mispronunciation Detection and Diagnosis*				
		Correct	*Accuracy*	*Precision*	*Recall*	*F-measure*	*Detection Accuracy*	*Diagnosis Accuracy*
Small Scale	APM	79.60%	72.20%	52.02%	84.67%	64.44%	84.24%	57.07%
	MT-APM	84.40%	76.80%	59.31%	**89.33%**	71.29%	87.86%	**75.69%**
	R-MT-APM	**86.10%**	**77.00%**	**63.47%**	88.78%	**74.02%**	**89.44%**	74.07%
	A-MT-APM	74.80%	61.80%	52.02%	84.67%	64.44%	84.24%	57.07%
Medium Scale	APM	80.80%	78.60%	53.22%	83.56%	65.02%	84.92%	73.47%
	MT-APM	85.50%	81.70%	61.29%	86.14%	71.62%	88.54%	75.83%
	R-MT-APM	**87.50%**	**83.10%**	**65.84%**	89.92%	**76.02%**	**90.44%**	**77.71%**
	A-MT-APM	83.70%	78.70%	62.26%	**90.35%**	73.72%	89.10%	73.77%
Large Scale	APM	81.40%	76.30%	63.35%	83.74%	72.13%	89.03%	68.36%
	MT-APM	86.40%	80.50%	62.78%	89.05%	73.64%	89.26%	**79.63%**
	R-MT-APM	**88.20%**	**83.30%**	67.65%	**89.52%**	**77.07%**	**90.99%**	78.24%
	A-MT-APM	86.80%	81.30%	**67.75%**	85.60%	75.63%	90.70%	75.72%

Table 2. The definitions in hierarchical evaluation

For all Phonemic Units		Recognition Result	
		Correct Pronunciation	Mispronunciation
Manually Transcribed Phonemic Unit	Correct Pronunciation	**TA**	**FR**
	Mispronunciation	**FA**	**TR (CD/DE)**

speakers (CHLOE-C) and 110 Mandarin speakers (CHLOE-M) [24]. 30% of each speaker audios are labeled by skilled linguists with L2 speakers' actual pronunciations (i.e. *Annotation* in Fig. 2). To eliminate the influence of inherent differences between Mandarin and Cantonese speakers, all our experiments are conducted only on CHLOE-C.

4.2. Experimental Setup

To evaluate the performance of models, three scales (small, medium and large) of CHLOE-C dataset are used. The data of 40, 60 and 80 speakers are randomly selected as training set, amounting to 5.0, 7.5 and 9.5 hours respectively. The data of another 20 speakers (2.0 hours) are selected as test set and keep consistent across experiments of different scales.

11 frames (5 before, 1 current and 5 after) of MFCC are used as the acoustic features x_t. 7 canonical phones (3 before, 1 current and 3after) are employed as the phonemic features (q_t^{Dict}).

Four different models are implemented for comparison: (1) baseline APM; (2) MT-APM; (3) R-MT-APM; and (4) A-MT-APM. The A-MT-APM is designed in a straightforward way where the CM-DNN outputs $\mathrm{P}(C|x_t,q_t^{Dict})$, $\mathrm{P}(M|x_t,q_t^{Dict})$ are directly appended to (x_t, q_t^{Dict}) as the input, and the other parts are the same as MT-APM. After preliminary experiments on network configurations, all the models have 7 hidden layers with 2048 units per layer and tanh as activation function. The CM-DNN contains 5 hidden layers with 512 units per layer and sigmoid as activation function.

4.3. Experimental Results

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

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

where N is the total number of labels, and S, D, I are the counts of substitution, deletion and insertion errors. The performance of MDD is evaluated with the hierarchical evaluation structure proposed in [25]. True Acceptance (TA), True Rejection (TR), False Rejection (FR), False Acceptance (FA), Correct Diagnosis (CD) and Diagnosis Error (DE) are defined as Table 2. Precision, Recall, F-Measure and accuracies of mispronunciation detection and diagnosis are used for performance measurement of MDD:

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

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

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

$$DiagnosisAccuracy\ = \frac{CD}{CD + DE} \qquad (11d)$$

The results are shown in Table 1. MT-APM and R-MT-APM approaches outperform baseline on both performance of phone recognition and MDD. Comparing R-MT-APM with APM shows significant F-Measure improvement from 64.44% to 74.02% (small scale), from 65.02% to 76.02% (medium scale) and from 72.13% to 77.07% (large scale).

5. CONCLUSION

This paper incorporates the multi-task learning technique into APM. The phone state posterior probabilities are derived in multi-task manner considering both correct recognition and mispronunciation recognition tasks. A feature representation module is then proposed to improve performance. Compared with the baseline APM, the proposed MT-APM and R-MT-APM achieve better performance. Through the representation module of input features, R-MT-APM achieves the highest mispronunciation detection accuracy.

6. ACKNOWLEDGEMENTS

This work is supported by a grant from the HKSAR RGC General Research Fund (project no. 14207315). The research was conducted while the first author was an intern at CUHK.

7. 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.", Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. Vol. 4, pp. IV-201, 2007.

[7] Zhang, F., Huang, C., Soong, F. K., Chu, M., and Wang, R., "Automatic mispronunciation detection for Mandarin.", Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on. IEEE, pp. 5077-5080, 2008.

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

[9] Strik, H., Truong, K., De Wet, F., and Cucchiarini, C., "Comparing different approaches for automatic pronunciation error detection.", Speech communication51.10 (2009), pp.845-852, 2009.

[10] 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.

[11] Qian, X., Meng, H., and Soong, F., "A two-pass framework of mispronunciation detection and diagnosis for computer-aided pronunciation training.", IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 24.6 (2016), pp.1020-1028, 2016

[12] 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.

[13] 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.

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

[15] 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.

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

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

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

[19] Seltzer, M. L., and Droppo, J., "Multi-task learning in deep neural networks for improved phoneme recognition.", Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, pp. 6965-6969, 2013.

[20] Chen, D., Mak, B., Leung, C. C., and Sivadas, S., "Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition.", Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, pp. 5592-5596, 2014.

[21] Parveen, S., and Green, P., "Multitask learning in connectionist robust ASR using recurrent neural networks", Eighth European Conference on Speech Communication and Technology, 2003.

[22] Duan, R., Kawahara, T., Dantsujii, M., and Zhang, J., "Pronunciation error detection using DNN articulatory model based on multi-lingual and multi-task learning", Chinese Spoken Language Processing (ISCSLP), 2016 10th International Symposium on. IEEE, pp.1-5, 2016

[23] Tong R, Chen N F, Ma B., "Multi-Task Learning for Mispronunciation Detection on Singapore Children's Mandarin Speech", Proc. Interspeech 2017, pp.2193-2197, 2017

[24] Meng, H., Lo, W. K., Harrison, A. M., Lee, P., Wong, K. H., Leung, W. K., and Meng, F., "Development of automatic speech recognition and synthesis technologies to support Chinese learners of English: The CUHK experience." Proc. APSIPA ASC, pp.811-820., 2010

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