CNN-RNN-CTC BASED END-TO-END MISPRONUNCIATION DETECTION AND DIAGNOSIS

Wai-Kim Leung, Xunying Liu and Helen Meng

Human-Computer Communications Laboratory,
Department of System Engineering and Engineering Management, The Chinese University of Hong Kong
Big Data Decision Analytics (BDDA) Research Centre, The Chinese University of Hong Kong

{wkleung, xyliu, hmmeng}@se.cuhk.edu.hk

ABSTRACT

This paper focuses on using Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Connectionist Temporal Classification (CTC) to build an end-to-end speech recognition for Mispronunciation Detection and Diagnosis (MDD) task. Our approach is end-to-end models, while phonemic or graphemic information, or forced alignment between different linguistic units, are not required. We conduct experiments that compare the proposed CNN-RNN-CTC approach with alternative mispronunciation detection and diagnoses (MDD) approaches. The F-measure of our approach is 74.65%, which significantly outperforms the Extended Recognition Network (ERN) (S-AM) by 44.75% and State-level Acoustic Model (S-AM) by 32.28% relatively. The relative improvement in F-measure when over Acoustic-Phonemic Model (APM), Acoustic-Graphemic Model (AGM) and Acoustic-Phonemic-Graphemic Model (APGM) are 9.57%, 5.04% and 2.77% respectively.

Index Terms— Computer Assisted Pronunciation Training (CAPT), Mispronunciation Detection and Diagnosis (MDD), Connectionist Temporal Classification (CTC), Convolutional Neural Network (CNN), e-learning

1. INTRODUCTION

Computer-assisted Pronunciation Training (CAPT) offers new opportunities for the language learning since the automated system is easily make available 24x7 for immersive learning, and can also solve the problem of teacher shortage. One of the key technologies of CAPT is the phone-level mispronunciation detection and diagnosis (MDD). MDD is more difficult than automatic speech recognition (ASR) since ASR can use the language model to outweigh the effect of inaccurate acoustics to output the legitimate character sequence. However, for MDD, the constraints offered by the language model is not helpful as it will lead to missed detection of mispronunciations. Hence, strong acoustic modelling is important in order to enable discriminate between native productions with canonical phonetic pronunciations and the deviant non-native pronunciations. This attracted wide research interest in recent years [1–7].

Mispronunciations in non-native productions may be attributed to language transfer of features from the primary language [3]. Phonological rules [3] are derived to model the mispronunciation patterns produced by learners. The phonological rules are applied to Extended Recognition Network (ERN) in traditional speech recognizer to enable the capability of MDD [7]. A CAPT online system [4, 5] had been developed with ERN to serve university students. However, ERN cannot guarantee that all mispronunciation possibilities from all language learners are covered. When the recognition network becomes overly bushy and covers too many mispronunciations, the acoustic model may not provide sufficient discrimination among the many alternative and the detection performance may drop.

In order to overcome the limitation of ERN, free-phone recognition was introduced. The state-level acoustic model was developed as the baseline approach for MDD [6]. Furthermore, the Acoustic-Phonemic Model (APM), Acoustic-Graphemic Model (AGM) and Acoustic-Phonemic-Graphemic Model (APGM) were also introduced for enhanced performances over the baseline. With the assistance of phoneme and grapheme information, the F-measure for MDD is increased from 51.55% (ERN (S-AM)) to 72.61% (APGM). However, forced-alignment is involved in these approaches. The accuracy of force-alignment would be the key factor of the performance of MDD. This leads us to investigate the approach without requirement of force-alignment for free-phone recognition in MDD.

Connectionist temporal classification (CTC) [8] was introduced to train the Recurrent Neural Network (RNN) for labelling unsegmented sequences directly. CTC has already been used in end-to-end speech recognition with acoustics-to-letter model [9] and acoustics-to-word model [10, 11]. Convolutional Neural Network (CNN) is also widely used in im-

age recognition tasks and have been applied effectively to many ASR tasks [12, 13]. Applying CNN to speech recognition model can reduce the phone error rate [14] and even make the model works better under mismatched (noisy) conditions [15].

In this paper, we propose to build the CNN-RNN-CTC model for MDD problem. Our model architecture is presented in Section 2. In Section 3, two experiments are conducted to evaluate the performance of the model. The performance of different hidden unit sizes will be discussed. The experimental results are also compared with other MDD approaches which are reported in [6]. Finally, the conclusion is presented in Section 4.

2. CNN-RNN-CTC MODEL FOR MDD

Our proposed CNN-RNN-CTC model consists of 5 parts and the architecture is shown in Figure 1. The first part is the input layer, it accepts the framewise acoustic features. This followed by a *batch normalization layer* and *zero padding layer*. The reason of adding zero padding layer is to ensure all utterances in a batch having the same length. The second part is convolution, it contains total 4 CNN layers, 2 Max-pool layers and followed by the batch normalization layer. The layer should capture the high level acoustic features from input layer. The third part is bi-directional RNN which is used to capture the temporal acoustic features. We use Gated Recurrent Unit (GRU) instead of Long Short-Term Memory (LSTM) since GRU is simpler than LSTM and it can speed up the training process. The fourth part is MLP layers (Time Distributed Dense layers), it ends with a softmax layer for the classification output. The last part is CTC output layer which is used to generate the predicted phoneme sequence.

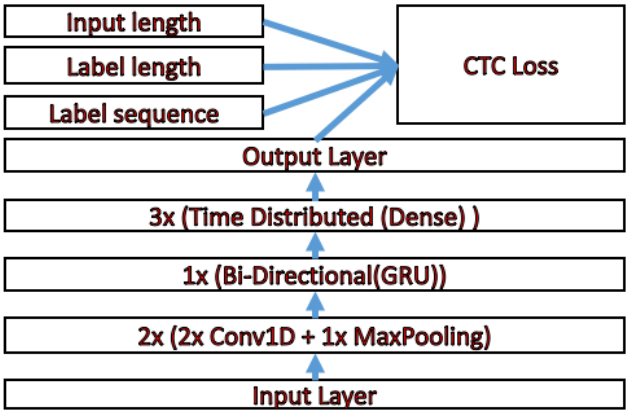

Fig. 1. The proposed CNN-RNN-CTC model

Table 1. Performance of phone recognition with different number of hidden units.

Hidden Size	Correct	Insertion	Deletion	Substitution
128	82.13% (85886)	2.67% (2791)	4.49% (4698)	10.71% (11199)
256	84.28% (88409)	2.98% (3129)	3.46% (3630)	9.28% (9733)
512	86.12% (90081)	2.72% (2850)	3.08% (3219)	8.08% (9449)
1024	**87.93%** (91877)	**2.15%** (2251)	**2.90%** (3033)	**7.01%** (7330)

3. EXPERIMENTS

3.1. Setup

3.1.1. Speech Corpus

We use TIMIT and CU-CHLOE (Chinese University CHinese Learners of English) corpora [16] to evaluate the performance of the our model. Both corpora are sampled with 16KHz, mono channels and recorded in sound proof room with close talking microphone. CU-CHLOE corpus includes 100 Cantonese speakers (50 males and 50 females) and 110 Mandarin speakers (60 males and 50 females). The content of the corpus includes: 1) the Aesops Fable The North Wind and the Sun, which has 6 sentences and covers all the English phonemes. 2) a set of 20 phonemic sentences designed by English teachers to cover common English mispronunciations. 3) a set of 10 pairs of confusing words from the TIMIT. 4) a set of 50 pairs of minimal pairs from the TIMIT. Every utterance is phonetically labeled by trained linguists.

All TIMIT data are grouped as training set. For CU-CHLOE, we split the corpus into training set, development set and test set following [6].

In this paper, two experiments are conducted to evaluate the model performance of phone recognition and MDD.

3.1.2. Model Training

The spectrogram is used as the input feature of the model. The FFT window size, step time, maximum frequency and hop size of spectrogram generation are set to 20ms,10ms, 8KHz and 160 respectively. Tensorflow and Keras are used to implement the model since it includes the CTC loss function. The input labels (annotated label sequence), the label length, input length and the softmax output from model are passed to CTC loss function to compute the loss. Experiments are conducted to investigate how the hidden unit size affects the performance with same number of layers. Each set of parameters are run for 30 epochs. The results with the highest F-measure are selected as the final result of the model.

Table 2. Performance of phone recognition with different approaches.

	Correct	Accuracy
CNN-RNN-CTC	90.08% (N.A.)	87.93% (N.A.)
S-AM [6]	81.15% (11.00%)	74.37% (18.23%)
ERN (S-AM) [6]	87.02% (3.52%)	83.17% (5.72%)
APM [6]	90.86% (-0.86%)	87.96% (-0.03%)
AGM [6]	91.14% (-1.16%)	88.74% (-0.91%)
APGM [6]	91.47% (-1.52%)	88.23% (-0.34%)

Note: Percentage in brackets indicate the relative percentage different when comparing to our CNN-RNN-CTC approach. Positive number means our model perform better.

Table 3. Confusion matrices of most frequency misrecognized vowels

		Annotation					
		aa	ah	ae	eh	ih	iy
CNN-RNN-CTC	aa	4874	286	43	3	3	0
	ah	215	**5321**	86	48	138	28
	ae	73	147	2155	228	16	3
	eh	6	52	129	1260	31	5
	ih	0	141	10	61	**3970**	232
	iy	1	25	2	19	216	2133
S-AM [6]	aa	4612	584	91	2	5	0
	ah	299	4280	208	104	476	33
	ae	68	258	1776	476	58	4
	eh	10	84	288	817	84	10
	ih	0	129	25	60	2637	382
	iy	0	29	7	19	613	1772
AGPM [6]	aa	**4929**	262	64	5	1	0
	ah	176	5113	73	21	194	54
	ae	65	160	**2304**	98	3	0
	eh	6	32	81	**1324**	31	5
	ih	1	51	6	28	3870	166
	iy	0	18	0	13	196	**2160**

Table 4. Confusion matrices of most frequency misrecognized consonants

		Annotation					
		d	dh	t	sh	s	z
CNN-RNN-CTC	d	**2420**	72	92	0	2	10
	dh	137	1448	9	0	5	38
	t	160	12	**7833**	2	30	12
	sh	1	1	1	844	50	2
	s	2	18	27	68	**4822**	146
	z	15	27	14	7	91	1160
S-AM [6]	d	1790	308	199	0	1	14
	dh	151	1142	59	1	11	68
	t	317	138	7082	2	31	38
	sh	0	1	12	814	89	2
	s	4	75	71	84	4336	542
	z	2	66	31	4	195	622
AGPM [6]	d	2332	28	86	0	2	4
	dh	204	**1917**	4	0	2	16
	t	93	9	7725	0	17	7
	sh	0	0	1	**850**	34	0
	s	0	7	35	59	4622	65
	z	1	2	5	2	113	**1303**

3.2. Evaluation

The recognized phone sequence are aligned with the annotated phone sequences (labelled by trained linguists) by using Needleman-Wunsch Algorithm [17]. The aligned phone sequences are used to count the correctness, insertion, deletion and substitution of the model. The results are presented in the following subsections.

3.2.1. Performance of Phone Recognition

The phone recognition performance is evaluated by aligning the annotated phone sequence and recognized phone sequence by the model. The experimental results of phone recognition are shown in Table 1. The best recognition result occurs when the size of hidden unit is 1024 and the correct phone recognition rate, insertion rate, deletion rate and substitution rate are 87.93%, 2.15%, 2.90% and 7.01% respectively.

The result is further compared with those reported in [6]. Table 2 shows the comparison result. The formulae for evaluation are as follows [18]:

$$Correct\ rate = \frac{N - S - D}{N} \tag{1a}$$

$$Accuracy = \frac{N - S - D - I}{N} \tag{1b}$$

where N is total number of labelled phones, and *S*, *D* and *I* represent for the count of substitution, deletion and insertion errors.

When only comparing the baseline model such as S-AM and ERN (S-AM) reported in [6], our model obtains the best phone correctness and accuracies (90.08% and 87.93% respectively), which are 11.00% and 18.23% relative better than S-AM and 3.52% and 5.72% relative better than ERN (S-AM). The results are further compared with APM, AGM and APGM and shown in Table 2. In the above three approaches, canonical phonetic information is present and it leads to higher recognition rate for correct pronunciation. As our proposed model lacks of phonemic and graphemic information, the correctness and accuracy slightly underperform these three approaches.

The confusion matrices of most frequently misrecognized vowels and consonants by S-AM and AGPM are shown in Table 3 and Table 4. Same experiment is conducted with our CNN-RNN-CTC model and shown in the same table. Our model significantly obtains better results than S-AM for all confusable consonants and vowels. The proposed CNN-RNN-CTC model has fewer confusions overall when compared with S-AM, but slightly underperforms when compared with AGPM, except for the vowels */ah/*, */ih/* and consonants consonants */d/*, */t/* and */s/*.

Table 5. Performance of MDD with different hidden unit

Hidden Unit	FRR	FAR	DER	Precision	Recall	F-measure	Detection Accuracy	Diagnosis Accuracy
128	16.00%	**15.76%**	23.49%	55.46%	**84.24%**	66.89%	84.05%	76.51%
256	13.80%	17.15%	21.70%	58.64%	82.85%	68.68%	85.56%	78.30%
512	10.75%	18.75%	19.40%	64.25%	81.25%	71.75%	87.71%	80.60%
1024	**8.66%**	18.85%	**16.76%**	**69.06%**	81.15%	**74.62%**	**89.38%**	**83.24%**

Table 6. Performance of MDD with different approach

	FRR	FAR	DER	Precision	Recall	F-measure	Detection Accuracy	Diagnosis Accuracy
CNN-RNN-CTC(AM)	8.66%	18.85%	16.76%	69.06%	81.15%	**74.62%**	89.38%	83.24%
ERN (S-AM) [6]	11.04%	43.59%	32.26%	42.39%	56.41%	51.55%	84.07%	67.74%
S-AM [6]	22.44%	15.70%	23.33%	42.39%	84.30%	56.41%	78.66%	76.67%
APM [6]	4.75%	36.61%	15.26%	73.57%	63.39%	68.10%	89.75%	84.74%
AGM [6]	5.25%	31.31%	13.49%	73.55%	68.69%	71.04%	90.18%	85.35%
APGM [6]	4.57%	30.53%	13.49%	76.05%	69.47%	72.61%	90.94%	86.51%

3.2.2. Performance of MDD

To evaluate the performance of MDD, we follow the hierarchical evaluation structure proposed in [19]. The true acceptance (TA) and true rejection (TR) rates indicate correct pronunciation detection, while False Reject (FR) and False Acceptance (FA) indicate incorrect detection. For the mispronunciation diagnosis, TR is further divided into Correct Diagnosis (CD) and Diagnosis Error (DE). The False Rejection Rate (FRR), False Acceptance Rate (FAR) and Diagnosis Error Rate (DER) are calculated by Equation 2:

$$FRR = \frac{FR}{TA + FR} \tag{2a}$$

$$FAR = \frac{FA}{FA + TR} \tag{2b}$$

$$DER = \frac{DE}{CD + DE} \tag{2c}$$

The evaluation measures of Precision, Recall and F-measure are widely used for measuring performance of MDD. The equations are defined as Equation 3.

$$Precision = \frac{TR}{TR + FR} \tag{3a}$$

$$Recall = \frac{TR}{TR + FA} = 1 - FAR \tag{3b}$$

$$F - measure = 2\frac{Precision * Recall}{Precision + Recall} \tag{3c}$$

For the accuracies of mispronunciation detection and mispronunciation diagnosis, the formulae are shown in Equations (4a) and (4b).

$$Detection\ Accuracy = \frac{TA + TR}{TA + FR + FA + TR} \tag{4a}$$

$$Diagnosis\ Accuracy = \frac{CD}{CD + DE} = 1 - DER \tag{4b}$$

Table 5 shows the performance of our model with different hidden unit sizes of MDD. The best F-Measure (74.62%) occurs when the size of hidden unit is 1024, we select this result and further compare with other approaches reported in [6].

Table 6 shows the performance of the MDD task of using our approach and the approaches reported in [6]. If we compare with the performance of ERN (S-AM), the relative improvements in F-measure, mispronunciation detection accuracy and mispronunciation diagnosis accuracy are 32.28%, 13.63% and 8.57% respectively. Also as shown in Table 6, although the proposed model did not use any phonemic and graphemic information, the relative improvement in F-measure when over APM, AGM and APGM are 9.57%, 5.04% and 2.77% respectively .

4. CONCLUSION

This paper presents the CNN-RNN-CTC approach to develop an end-to-end speech recognition approach for the task of MDD, such that it does not require any explicit phonemic and graphemic information input and hence no forced alignment is required. Our approach do not need the presence of any phonemic and graphemic information and no force-alignment is required. The experiment results show that our approach significantly outperform previously approaches, including those that utilize phonemic and graphemic information input, by a relative improvement of 9.57% over APM, 5.04% over AGM and 2.77% over APGM. This work can be used as the baseline of end-to-end approach for the task of MDD. In the future, we will work on adding the linguistic information to further improve the performance.

5. ACKNOWLEDGEMENTS

This work is partially supported by the grant from the HKSAR Government Research Grants Council General Research Fund (project number 14207315). We thank Dr. Li Kun and Mr. Li Xu for providing the dataset reported in [6].

6. REFERENCES

[1] 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.

[2] Wei Li, Sabato Marco Siniscalchi, Nancy F Chen, and Chin-Hui Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with dnn-based speech attribute modeling," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 6135–6139.

[3] Wai-Kit Lo, Shuang Zhang, and Helen Meng, "Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[4] Ka-Wa Yuen, Wai-Kim Leung, Peng-fei Liu, Ka-Ho Wong, Xiao-jun Qian, Wai-Kit Lo, and Helen Meng, "Enunciate: An internet-accessible computer-aided pronunciation training system and related user evaluations," in *Speech Database and Assessments (Oriental COCOSDA), 2011 International Conference on*. IEEE, 2011, pp. 85–90.

[5] Pengfei Liu, Ka-Wa Yuen, Wai-Kim Leung, and Helen Meng, "menunciate: Development of a computer-aided pronunciation training system on a cross-platform framework for mobile, speech-enabled application development," in *Chinese Spoken Language Processing (ISCSLP), 2012 8th International Symposium on*. IEEE, 2012, pp. 170–173.

[6] Kun Li, Xiaojun Qian, and Helen Meng, "Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 193–207, 2017.

[7] Alissa M Harrison, Wai-Kit Lo, Xiao-jun Qian, and Helen Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training," in *International Workshop on Speech and Language Technology in Education*, 2009.

[8] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[9] Alex Graves and Navdeep Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning*, 2014, pp. 1764–1772.

[10] Kartik Audhkhasi, Bhuvana Ramabhadran, George Saon, Michael Picheny, and David Nahamoo, "Direct acoustics-to-word models for english conversational speech recognition," *arXiv preprint arXiv:1703.07754*, 2017.

[11] Kartik Audhkhasi, Brian Kingsbury, Bhuvana Ramabhadran, George Saon, and Michael Picheny, "Building competitive direct acoustics-to-word models for english conversational speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4759–4763.

[12] Ying Zhang, Mohammad Pezeshki, Philémon Brakel, Saizheng Zhang, Cesar Laurent Yoshua Bengio, and Aaron Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," *arXiv preprint arXiv:1701.02720*, 2017.

[13] William Chan and Ian Lane, "Deep convolutional neural networks for acoustic modeling in low resource languages," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2056–2060.

[14] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, and Gerald Penn, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4277–4280.

[15] Dimitri Palaz, Ronan Collobert, et al., "Analysis of cnn-based speech recognition system using raw speech as input," in *Proceedings of INTERSPEECH*, 2015, number EPFL-CONF-210029.

[16] Helen Meng, Yuen Yee Lo, Lan Wang, and Wing Yiu Lau, "Deriving salient learners mispronunciations from cross-language phonological comparisons," in *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*. IEEE, 2007, pp. 437–442.

[17] Vladimir Likic, "The needleman-wunsch algorithm for sequence alignment," *Lecture given at the 7th Melbourne Bioinformatics Course, Bi021 Molecular Science and Biotechnology Institute, University of Melbourne*, pp. 1–46, 2008.

[18] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al., "The htk book," *Cambridge university engineering department*, vol. 3, pp. 175, 2002.

[19] Xiaojun Qian, Helen Meng, and Frank Soong, "Capturing l2 segmental mispronunciations with joint-sequence models in computer-aided pronunciation training (capt)," in *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*. IEEE, 2010, pp. 84–88.