INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Improving Mispronunciation Detection for Non-Native Learners with Multisource Information and LSTM-Based Deep Models

Wei Li[1], Nancy F. Chen[2], Sabato Marco Siniscalchi[1, 3], and Chin-Hui Lee[1]

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, USA
[2]Institute for Infocomm Research, Singapore
[3]Department of Telematics, Kore University of Enna, Enna, Italy

{lee.wei, chl}@gatech.edu, marco.siniscalchi@unikore.it, nfychen@i2r.a-star.edu.sg

Abstract

In this paper, we utilize manner and place of articulation features and deep neural network models (DNNs) with long short-term memory (LSTM) to improve the detection performance of phonetic mispronunciations produced by second language learners. First, we show that speech attribute scores are complementary to conventional phone scores, so they can be concatenated as features to improve a baseline system based only on phone information. Next, pronunciation representation, usually calculated by frame-level averaging in a DNN, is now learned by LSTM, which directly uses sequential context information to embed a sequence of pronunciation scores into a pronunciation vector to improve the performance of subsequent mispronunciation detectors. Finally, when both proposed techniques are incorporated into the baseline phone-based GOP (goodness of pronunciation) classifier system trained on the same data, the integrated system reduces the false acceptance rate (FAR) and false rejection rate (FRR) by 37.90% and 38.44% (relative), respectively, from the baseline system.

Index Terms: mispronunciation detection and diagnosis, computer assisted pronunciation training (CAPT), automatic speech attribute transcription (ASAT), deep neural network (DNN), Long Short-Term Memory (LSTM)

1. Introduction

With accelerating globalization, more and more people are willing or required to learn second languages (L2). Among them, Mandarin is becoming increasingly popular with at least 100 million people taking Mandarin as their second language [1, 2, 3]. Meanwhile, computer assisted language learning (CALL) [4] systems can play a key role in alleviating the lack of qualified teachers and offering flexibility in terms of time and space constraints. It is well-known that the L2 learning process is heavily affected by a well-established habitual perception of phonemes and articulatory motions in the learners' primary language (L1) [5], which often cause mistakes and imprecise articulation in speech productions by the L2 learners, e.g., a negative language transfer [5, 6]. Therefore, computer assisted pronunciation training (CAPT), a critical component of a CALL system, is often employed to automatically assess L2 learners' pronunciation quality, detect mispronunciations and provide corrective feedbacks.

A wealth of research work has utilized confidence scores [7-12] derived from automatic speech recognition (ASR) systems to provide pronunciation scores to the L2 learners. For example, the log-likelihood ratio (LLR) was adopted in [7] as a confidence score to measure the difference between native and non-native acoustic phone models. A variation of the posterior probability ratio, a "Goodness of Pronunciation (GOP)" [8] score, was also proposed to evaluate the L2 learners' pronunciation quality. Along with its variations [9, 10], GOP score has been widely used to detect mispronunciations as well. After formulating mispronunciation detection as a binary classification task, GOP scores between a canonical phone and other competing phones are combined into a feature vector, which is then fed into phone-dependent classifiers (e.g., [10, 11]). The posterior probabilities obtained from those classifiers are often used as pronunciation scores. However, when facing lower confidence scores, L2 learners are more likely to feel helpless, because they do not know what is wrong with their pronunciation and how to improve it when only given numeric scores. In [12], it was shown that L2 learners could improve their production of the targeted phones by receiving a corrective feedback about the mispronunciation error at the phone level. More recent research work has thus focused on how to use automatic mechanisms to generate finer detection results and corrective information, such as in the phone-based extended recognition network (ERN) [13, 14] approach. To reduce the amount of resources needed for collecting frequent L1-dependent error patterns, a recent work [15] proposed an acoustic phonological model to automatically learn the acoustic-phonetic rules from canonical productions of words and annotated mispronunciation. Consequently, a multi-distributed DNN [15] leveraging learned phonological rules is then used to accomplish phone recognition and provide phone-level corrective feedback.

Although the above-mentioned phone-based CAPT systems have achieved satisfactory mispronunciation detection results, the performance is often heavily dependent on the quality of phone-level labeling of the non-native corpora used for training the phone models for pronunciation scoring. Labeling non-native speech data is intrinsically much more challenging than labeling native speech data. In [16, 17], it was observed that L2 learners' mispronunciations contain many "distortion errors", i.e., the erroneous pronunciation is often between two canonical phones, rather than a straight-forward phonemic substitution. Therefore, standard forced-assigned human labeling of phone categories inevitably becomes noisy phone labels during acoustic model training. In addition, as phonetic annotation is a subjective task, even for linguistic experts, annotator subjectivity adds another layer of complexity to the ground-truth labels.

Faced with the challenges of inconsistency in non-native phone-based labeling and imperfect acoustic modeling, our previous work [18, 19] has investigated articulatory-based modeling for CAPT, where speech attributes [20, 21]

http://dx.doi.org/10.21437/Interspeech.2017-464

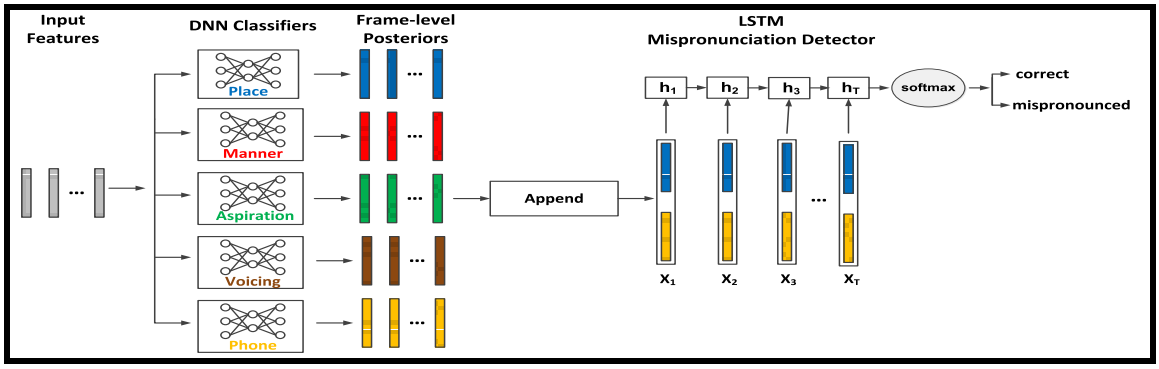

Figure 2: *Overview of mispronunciation detection framework*

describing articulatory characteristics are proposed to directly measure pronunciation quality and give corrective feedbacks based on articulation manner and place. Speech attribute features have been used as complementary features to reduce word error rate in ASR, e.g., [22, 23, 24], where articulatory-motivated features are shown to help improve robustness toward noise, speaking styles and speaker population. In addition, the phone sharing mechanism that each speech attribute is shared by a group of phones, results in each speech attribute leveraging more training data from a group of phones, so that the side effect of labeling noise at the individual phone-level is mitigated. In this work, we combine attribute and traditional phone features to improve the performance of phone-level mispronunciation detection. In addition, traditional frame-level averaging feed-forward neural network based classifier [11] would be replaced by an LSTM [25], which uses learned sequential context information to embed a sequence of pronunciation scores into a pronunciation vector. Compared with the pronunciation representation derived from traditional frame-level averaging [11, 18], LSTM embedded pronunciation vector is expected to contain much more context/historical information for subsequent mispronunciation detection.

2. Mandarin Phones & Speech Attributes

We focus on learners of Mandarin Chinese with European first languages. Each Chinese character corresponds to one spoken syllable, consisting of an initial, usually a consonant, and a final, usually a vowel or sometimes a vowel followed by a nasal consonant. There is a total of 21 syllable initials and 38 syllable finals. As a preliminary study, we are concerned with mispronunciation of 21 syllable initials, because initial errors are more prone to cause miscommunication in Mandarin when compared to finals [26]. Moreover, after analyzing a large-scale non-native Mandarin corpus, Chen [27] found that 90% of top 10 mispronounced phones are syllable initials.

Each initial's articulatory characteristic can be described using its corresponding speech attribute [20, 21, 28 29]. In addition to manner and place of articulation, we also consider voicing and aspiration. Voicing is used to describe if the vocal cords vibrates; whereas, aspiration is used to describe whether there is a brief puff of air after an obstruction is released. For example, speech attribute features "labial", "unaspirated" and "stop" are used to describe how the initial "B" is produced. The detailed mapping table between speech attribute features and Mandarin initials can be found in our previous work [18, 19].

To visualize the degree of phone label inconsistency, we compared the histogram of phone-dependent segmental posterior of native and non-native corpus in Figure 1, where the human labeling phone category is Mandarin initial /J/, one of the top mispronounced Mandarin phones produced by L2 learners [27, 30]. More canonical pronunciations from native speakers are expected to have higher posteriors close to 1, i.e., the histogram on the upper panel of Figure 1. For non-native speakers, we see the posterior scores are more evenly spread out, i.e., the subfigure on the lower panel of Figure 1. The above-mentioned segmental posterior is achieved by averaging frame-level phone posteriors calculated using Eq. (1).

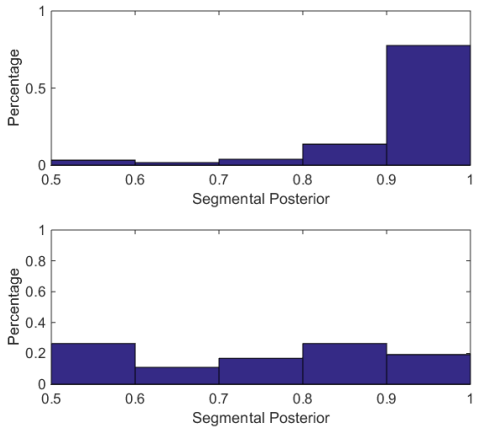

Figure 1: *Segmental posterior histograms of Mandarin Initial /J/ computed on native (upper) and non-native data (lower)*

3. Overview of Detection Framework

Figure 2 shows the proposed mispronunciation detection framework, which consists of two blocks: (i) the speech attribute and phone feature extraction module, which has also been used in the automatic speech attribute transcription (ASAT) paradigm [24], and (ii) the phone-dependent mispronunciation detector training module, which is based on a sequence of frame-level features.

3.1. Speech Attribute and Phone Feature Extraction

As in [24], speech attribute features are extracted using a bank of speech attribute detectors. A DNN-based classifier is separately trained for each articulatory-motivated attribute category described in [18, 19]. A window of 11 speech frames centered on the current frame is fed into each DNN classifier, which in turn generates a set of confidence scores in terms of posterior probabilities that the current frame pertains to each possible attribute within the target category. We name these

posteriors as frame-level speech attribute features. Similarly, a DNN phone classifier analyzes an expanded frame of the input speech signal and produces the posterior probability that pertains to each tied HMM state, often referred to as senone. Subsequently, [11, 31] proposed to use Eq. (1) to calculate the posteriors of current frame belonging to each phone category.

$$P(p|o_t) = \sum_{s \in p} P(s|o_t) \quad (1)$$

where unit p is the target phone category, o_t is the input feature at frame *t*, and s is the senone label used to compute phone state posterior; $\{s \in p\}$ is the set of all senones corresponding to unit p. Finally, a sequence of frame-level speech attribute and phone features (posteriors) are concatenated and fed to the LSTM mispronunciation detection module.

3.2. Mispronunciation Detector Construction

After receiving the extracted feature sequences and the phone-level time boundary information (obtained through forced-alignment), an LSTM is trained to learn the decision boundaries between the correct and mispronounced samples. LSTMs have been applied to solve many sequential classification tasks [32, 33 and 34] and achieved superior classification accuracy. Similar to sentiment analysis and text classification tasks [32, 34] in natural language processing, we first use LSTM to map the input sequence into a fixed-size feature vector (the output of the hidden layer at the last time step), which is then fed into a softmax layer for binary classification. The LSTM transition equations are defined as follows:

$$i_t = sigmoid(W_i x_t + U_i h_{t-1} + b_i) \quad (2)$$

$$f_t = sigmoid\big(W_f x_t + U_f h_{t-1} + b_f\big) \quad (3)$$

$$o_t = sigmoid(W_o x_t + U_o h_{t-1} + b_o) \quad (4)$$

$$g_t = tanh\big(W_g x_t + U_g h_{t-1} + b_g\big) \quad (5)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (6)$$

$$h_t = o_t \odot tanh(c_t) \quad (7)$$

where the operation $\odot$ denotes the element-wise vector product. At each time step *t*, x_t is the current input, i_t, f_t, o_t, g_t are gate functions defined in [25], c_t is memory cell and h_t is hidden layer representation. W_* and b_* denote the weight matrix and bias vector of corresponding gate functions.

Given a speech attribute and phone feature sequence $x = \{x_1, x_2 \cdots x_T\}$, we adopt Eq. (2) – (7) iteratively from t = 1 to T to calculate the hidden layer at the last time step h_T, which is fed into Eq. (8) to predict the probability distribution over pre-defined classes.

$$\hat{y} = softmax(W h_T + b) \quad (8)$$

where W and b is the weight matrix and bias vector of the non-linear softmax layer.

The above parameters are trained to minimize the cross-entropy function of the predicted and true distributions, as shown in Eq. (9).

$$L(y, \hat{y}) = -\sum_{i=1}^{N} \sum_{j=1}^{C} y_i^j \log\big(\hat{y}_i^j\big) \quad (9)$$

where N is the number of the training samples and C denotes the class number. y_i^j and $\hat{y}_i^j$ are the ground-truth and predicted probability distributions.

4. Experiments

4.1. Speech Corpora

Two speech corpora, (i) a native speech corpus from the Chinese National Hi-Tech Project 863 for Mandarin LVCSR system development [35], and (ii) a subset of a non-native speech corpus called iCALL [36], are mixed together to train our speech attribute and phone classifiers. The detailed description of our training and testing sets can be found in our previous work [18, 19].

4.2. Speech Attribute and Phone Feature Extraction Setup

The input feature (see Figure 2) is a window of 11 speech frames, each includes a 69-dimension of FBANK+Δ+ΔΔ vector. Phone and its senone labels were derived from forced-alignment with a GMM-HMM system. Speech attribute labels were next obtained using phone-attribute mapping table [18, 19]. We adopted those labels to separately train the set of corresponding DNNs. Each DNN has 6 hidden layers each with 2048 sigmoid units. The softmax function was employed at the output layer. The output dimension of each speech attribute classifier is the size of each category. The phone classifier's output dimension is 3487 (that is, the number of senones). After configuring the DNN architecture, we used the Kaldi toolkit [37] to train speech attribute and phone classifiers. At evaluation time, DNNs map the input feature vectors into frame-level speech attribute and phone senone posteriors. Finally, we compute frame-level phone posterior with Eq. (1).

4.3. LSTM Mispronunciation Detection Setup

Frame-level speech attribute and phone posteriors are first concatenated to form a new feature vector. A phone-dependent LSTM mispronunciation detector is then trained on this vector according to the phone level labels (correct/incorrect). Each phone's time boundary is obtained from forced-alignment. LSTMs are built with the KERAS toolkit [38]. The LSTM has two hidden layers each with 128 memory cells. The Adam optimizing algorithm [39] is chosen to minimize the cross-entropy described in Eq. (9). Before training the LSTM parameters, two data pre-processing steps have been executed to deal with variable length of input sequences and data imbalance problems. Zero-padding was performed to pad shorter consonant segment based on the maximum length in training set to meet the requirement of KERAS [38]. Regarding the data imbalance problem, namely the number of correct samples is higher than that of the incorrect samples leading to a biased mispronunciation detector with a high precision rate but a low recall rate, we used other phones' correctly pronounced samples as the target phone's incorrect samples, as in [19, 40].

4.4. Experimental Results and Discussion

Following previous work [18, 19, 40] on CALL system evaluation, two metrics, false acceptance rate (FAR) and false rejection rate (FRR) are used to measure the system performance. Under the conventional phone-based system framework, segmental phone scores are first calculated by frame-level averaging within each phone segment. Subsequently calculated phone segmental posterior (pronunciation vector) is fed into an ANN baseline to detect whether the current phone segment is mispronounced. Like previous work [11, 18], the ANN baseline has one hidden layer each with 512 non-linear activation nodes.

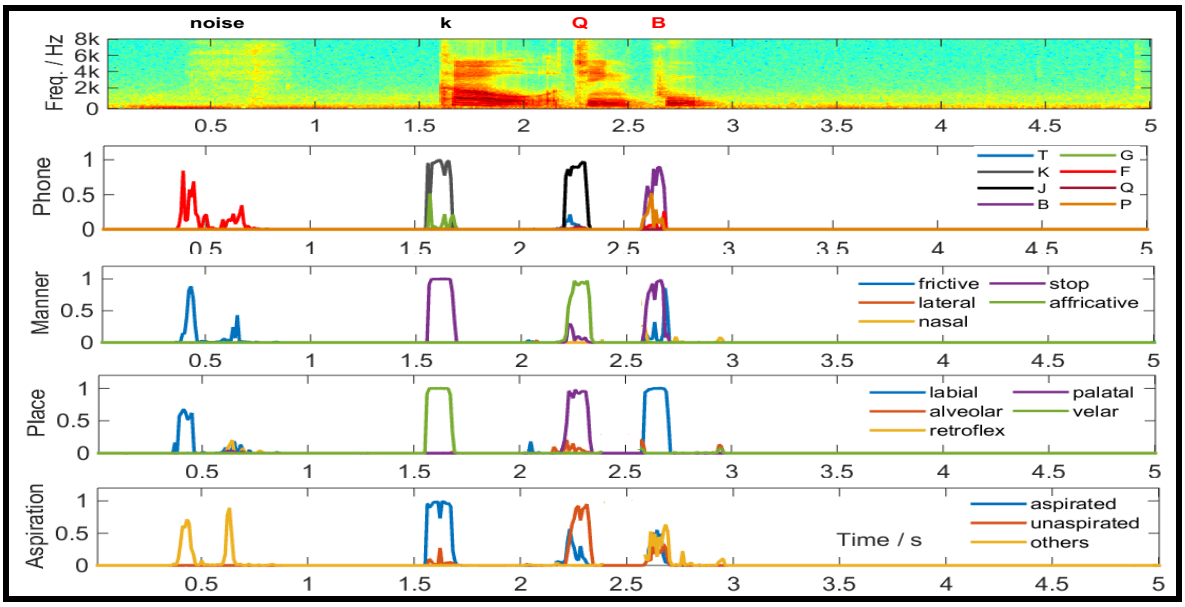

Figure 3: *Frame-level posteriors of each phonetic and articulatory category.*

After choosing the LSTM embedded vector as the segmental pronunciation representation for mispronunciation detection, the LSTM-phone-based system, shown in fourth row in Table 1, reduces FAR from 13.93% to 12.46%, and FRR from 5.02% to 3.87%, compared with ANN-phone-based baseline, shown in the second row, where direct frame-level averaging is used to calculate the pronunciation score vector. This observation confirms that the context information characterized and modeled by LSTM is helpful for verifying phone mispronunciations. With appended features, the LSTM system results, shown in the bottom row in Table 1, gives a further error reduction of FAR and FRR, implying that multisource information enables LSTM to learn a more accurate pronunciation representation for mispronunciation detectors.

Table 1. *Mispronunciation detection performance of four different systems on the test set*

Systems	FAR	FRR
ANN (phone features)	13.93	5.02
ANN (phone + attribute features)	12.61	4.87
LSTM (phone feature)	12.46	3.87
LSTM (phone + attribute features)	8.65	3.09

Mispronunciation systems trained with appended speech attribute features outperform systems trained with only conventional phonetic features, independently of using ANN- or LSTM-based mispronunciation detectors. Experimental results demonstrate that combining articulatory-motivated attribute and phone features is more robust to detect non-native speakers' mispronunciation. Figure 3 is an example of using multiple information to detect mispronunciations. In the upper panel, the spectrogram of three syllables and their corresponding canonical consonants are shown, where red color denotes mispronunciations. Detection curves of the aspirated affricative palatal phone /Q/ concurrently showed that it's mispronounced to its unaspirated counterpart phone /J/. A subset of phone detection curves in the second panel show that /B/ has the highest frame-level posterior, so its pronunciation is correct. However, this would result in a false acceptance, since its expected "unaspirated" score is low, as shown in the fifth panel. This demonstrates that articulatory information is complementary to traditional phone features. Previous research had already shown that articulatory-motivated features compensate for acoustic variations of native speakers and speaking styles [22-24]. The observed improvement might also be attributed to the sharing mechanism of speech attribute, i.e., each speech attribute feature is shared by a group of phones, which allows it to pool more training data than an individual phonetic category, so that the speech attribute features are not as sensitive to individual phone-level labeling errors *[1] and could be more robustly trained.

5. Conclusion

In this paper, speech attribute features are shown to be complementary to the conventional phone features. When merging them into pronunciation representations as inputs to the LSTM based classifiers to detect phone mispronunciations at the segment level, we show that the combined features are less sensitive to noisy phone-level labels of non-native corpora. Furthermore, when modeling dynamic changes of frame-level pronunciation scores, the proposed framework significantly reduces the FAR and FRR by 37.90% and 38.44% (relative), respectively, from the baseline system with phone features. For future work, the proposed system will be extended to detecting Mandarin vowel mispronunciation and giving corrective feedback based on articulator parameters, such as the tongue position and lip roundness. Furthermore, multi-way feedback visualization tools, such as information embedded in decision trees [19] and detection curves [18] will be investigated.

*Facing L2 learners' distorted pronunciation, where the non-native speech segment is between two canonical phones, the annotator is asked to assign a single phonetic category, as too overly detailed annotations will result in insufficient training data for engineering modeling purposes and make labeling consistency of L2 learners' mispronunciations much more challenging.

6. References

[1] CCTV, http://www.cctv-america.com/2015/03/03/chinese-as-a-second-language-growing-in-popularity, 2015.

[2] Kennedy S, Parker D A. Building China's 'One Belt, One Road'[J]. Center for Strategic and International Studies, 2015.

[3] Swaine M D. Chinese views and commentary on the 'One Belt, One Road 'initiative[J]. China Leadership Monitor, 2015.

[4] Gündüz, Nazlı. "Computer assisted language learning." Journal of Language and Linguistic Studies, 2005.

[5] Ellis, R. The Study of Second Language Acquisition. Oxford University Press. 1994.

[6] H. Meng, "Developing speech recognition and synthesis technologies to support computer-aided pronunciation training for Chinese learners of English," In *Proc. 23rd Pacific Asia Conference on Language, Information and Computation*, 2009.

[7] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, "Automatic detection of phone-level mispronunciation for language learning," in *Proc. Eurospeech*, 1999.

[8] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[9] J. Zheng, C. Huang, M. Chu, F. K. Soong, and W. Ye, "Generalized Segment Posterior Probability for Automatic Mandarin Pronunciation Evaluation," in *Proc. ICASSP*, 2007.

[10] S. Wei, G. Hu, Y. Hu, R.H. Wang, "A new method for mispronunciation detection using support vector machine based on pronunciation space models," *Speech Communications*, vol. 51, no. 10, pp. 896–905, 2009.

[11] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved Mispronunciation Detection with Deep Neural Network Trained Acoustic Models and Transfer Learning based Logistic Regression Classifiers," *Speech Communication*, 67, pp. 154- 166, 2015.

[12] A. Neri, C. Cucchiarini ,and H. Strik, "ASR-based corrective feedback on pronunciation: does it really work?," in Proc. Interspeech, 2006.

[13] H. Meng, Y. Lo, L. Wang, and W. Yiu, "Deriving salient learners' mispronunciations from cross-language phonological comparisons," in *Proc. ASRU*, 2007.

[14] W. K. Lo, S. Zhang and H. Meng, "Automatic derivation of phonological rules for mispronunciation detection in a Computer-Assisted pronunciation training system," in *Proc. Interspeech*, 2010.

[15] Li, Kun, 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, 2017.

[16] S. Yoon, M. Hasegawa-Johnson, and R. Sproat, "Landmark Based Automated Pronunciation Error Detection", in *Proc. Interspeech*, 2010.

[17] R. Duan, et al, "A Preliminary Study on ASR-based Detection of Chinese Mispronunciation by Japanese Learners," in *Proc. Interspeech*, 2014

[18] W. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, "Improving Non-native Misprounciation Detection And Enriching Diagnostic Feedback With DNN-BASED Speech Attribute Modeling", in *Proc. ICASSP*, 2016.

[19] W. Li, et al. "Detecting Mispronunciations of L2 Learners and Providing Corrective Feedback Using Knowledge-Guided and Data-Driven Decision Trees." in *Proc. Interspeech* 2016.

[20] K. N. Stevens, Acoustic Phonetics. Cambridge, MA, *MIT Press*, 2000.

[21] G. Fant, Speech Sounds and Features. Cambridge, MA, MIT Press, 1973.

[22] Kirchhoff, Katrin. "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," in *Proc. ICSLP*. 1998.

[23] Metze, Florian, and Alex Waibel. "A flexible stream architecture for ASR using articulatory features," in *Proc. Interspeech*. 2002.

[24] C.-H. Lee and S. M. Siniscalchi, "An information-extraction approach to speech processing: Analysis, detection, verification, and recognition," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1089– 1115, 2013.

[25] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory," Neural computation, 1997.

[26] J.-S. Zhang, W. Li, et al, "A Study On Functional Loads of Phonetic Contrasts Under Context Based On Mutual Information of Chinese Text And Phonemes," in *Proc. ISCSLP*, 2010.

[27] Chen, Nancy F., et al. "iCALL corpus: Mandarin Chinese spoken by non-native speakers of European descent," in *Proc. INTERSPEECH.* 2015.

[28] 林焘, 王理嘉, 语音学教程[M]. 北京大学出版社, 2013.

[29] 张家騄, 汉语人机语音通信基础[M]. 上海科学技术出版社, 2010.

[30] Chia-Yu Chiu, Yuan-Fu Lia, Daniel Kulls, Hansjorg Mixdorff, and Shing-Lung Chen, "A preliminary study on corpus design for computer-assisted German and Mandarin language learning," in *Proc COCOSDA*. 2009.

[31] W. Hu, Y. Qian, and F. K. Soong, "An improved DNN-based approach to mispronunciation detection and diagnosis of L2 learners' Speech," in *Proc. SLaTE*, 2015

[32] Liu, Pengfei, Xipeng Qiu, and Xuanjing Huang. "Recurrent neural network for text classification with multi-task learning." arXiv preprint arXiv:1605.05101 (2016).

[33] Chao, Linlin, et al. "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. ACM, 2015.

[34] Wang, Xingyou, Weijie Jiang, and Zhiyong Luo. "Combination of convolutional and recurrent neural network for sentiment analysis of short texts." Proceedings of the 26th International Conference on Computational Linguistics. 2016.

[35] S. Gao, *et al*, "Update of Progress of Sinohear: Advanced Mandarin LVCSR System At NLPR," in *Proc. ICSLP*, 2000.

[36] N. F. Chen et al., "Large-Scale Characterization of Non-Native Mandarin Chinese Spoken by Speakers of European Origin: An Analysis on iCALL," Speech Communication, 2016.

[37] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[38] P.W.D. Charles, "KERAS", GitHub repository, https://github.com/charlespwd/keras, 2013.

[39] Kingma, Diederik, and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980, 2014.

[40] J. Jiang and B. Xu, "Exploring the automatic mispronunciation detection of confusable phones for Mandarin," in *Proc. ICASSP*, 2009.