IMPROVING MANDARIN TONE MISPRONUNCIATION DETECTION FOR NON-NATIVE LEARNERS WITH SOFT-TARGET TONE LABELS AND BLSTM-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, GA. USA
[2]Institute for Infocomm Research, Singapore
[3]Department of Computer Engineering, Kore University of Enna, Enna, Italy
lee.wei@gatech.edu, nfychen@i2r.a-star.edu.sg, marco.siniscalchi@unikore.it, chl@ece.gatech.edu

ABSTRACT

We propose three techniques to improve mispronunciation detection of Mandarin tones of second language (L2) learners using tone-based extended recognition network (ERN). First, we extend our model from deep neural network (DNN) to bidirectional long-short-term memory (BLSTM) in order to model tone-level co-articulation influenced by a broader temporal context (e.g., two or three consecutive Mandarin syllables). Second, we relax the hard labels to characterize the situations when a single tone class label is not enough because L2 learners' pronunciations are often between two canonical tone categories. Therefore, soft targets (a probabilistic transcription) are proposed for acoustic model training in place of conventional hard targets (one-hot targets). Third, we average tone scores produced by BLSTM models trained with hard and soft targets to seek the complementarity from modeling at the tone-target levels. Compared to our previous system based on the DNN-trained ERNs, the BLSTM-trained system with soft targets reduces the equal error rate (ERR) from 5.77% to 4.86%, and system combination decreases EER further to 4.34%, achieving a 24.78% relative error reduction.

Index Terms— Computer assistant language learning (CALL), computer assisted pronunciation training (CAPT), tone recognition and mispronunciation detection, deep learning

1. INTRODUCTION

Computer assisted pronunciation training (CAPT) tools are becoming more and more useful for second language (L2) learners. Because these tools alleviate the lack of qualified teachers and offer flexibility in terms of time and space constraints. CAPT tools not only measures pronunciation quality at the segmental level (e.g., individual phonetic units [1-5]), but also mispronunciations at the supra-segmental level (e.g., stress [6-7], tone [8-18] and intonation [19-20]) could also be detected. Lexical tone error is a special type of mispronunciation often arising in tonal languages, such as Mandarin Chinese, in which each word is composed of one to several characters; each is pronounced as a basic syllable with five different tones including four lexical tones (see Figure 1) and one neutral tone (0). Different tones make the basic syllable have different lexical meanings, e.g., *ma1* (mother), *ma2* (hemp), *ma3* (horse) *and* ma4 (scold) have the same toneless syllable, namely *ma*, yet different tone markers. Obviously, tone mispronunciations are prone to cause miscommunication in Mandarin. Hence, tone mispronunciation detection subsystem [8-18] plays a key role in Mandarin CAPT systems.

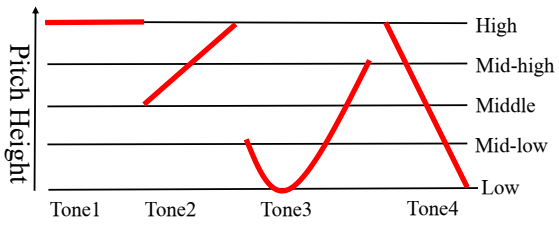

Figure 1: *Pitch contours of standard Mandarin lexicon tones*

Over the past two decades, many methods have been proposed to detect tone-level mispronunciations. The template-based detection framework was first introduced in [10], where discrete pitch value and duration of the tested tone is compared with the canonical tone template to decide whether the current tone is mispronounced. In [11], the Euclidean distance was adopted to calculate the distance between L2 learners' contours and ideal contours in energy and pitch domain. Although template-based approaches obtained acceptable performance, a superior mispronunciation detection performance was constantly achieved by using statistical model-based framework [12-16], in which pitch-related features (e.g., pitch, energy [12-15] or fundamental frequency variation [16]) are firstly extracted to train tone classifiers, e.g. decision tree [15], HMM [12-13], GMM [16] and ANN [14]. Subsequently, given test tone segment, its posterior or goodness of pronunciation (GOP) score measures the tone-level pronunciation quality. Inspired by recent findings on the role played by cepstral features in tone recognition within a DNN framework [21-22] along with the understanding that tone pitch realization and perception are influenced by the underlying phone units [23, 24], several authors have proposed DNNs to map combined cepstral and tonal features to tone-related posteriors [8, 9, 17, 18], which in turn are fed into classifiers or GOP calculators to verify the pronunciation correctness of the current tone.

Although the abovementioned model-based CAPT systems have achieved satisfactory mispronunciation detection results, there is still room for further improvement. Compared with phone-level co-articulation, the tone is influenced by a broader temporal context (e.g., 2 or 3 neighboring syllables) [25]. In [8], we have demonstrated that better results could be achieved by expanding the temporal information injected into the DNN-based system; nevertheless, the temporal information is still limited to the fixed context window spanning between 11 and 21 consecutive speech frames. Furthermore, L2 learners are more likely to slow down their pronunciation speed and increase each syllable's duration due to

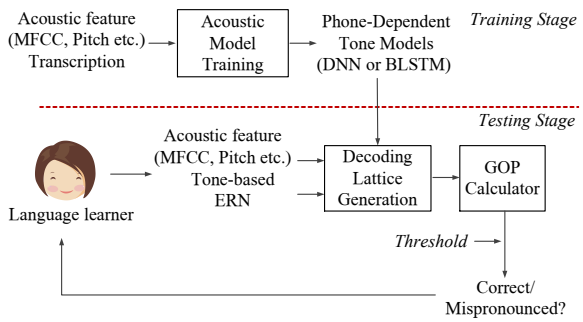

Figure 2: *A block diagram of the proposed mispronunciation detection system.*

their unfamiliarity with tone production, and that phenomenon is especially true for *tone2* and *tone3* [26-28]. In this paper, we replace DNN with bidirectional long-short-term memory (BLSTM) in [29] because the latter can handle bidirectional long-term dependencies of acoustic and prosodic features, and it can thereby better model tone-level co-articulation and irregular non-native tone production.

The performance of model-based CAPT system also heavily depends upon the quality of tone-level labeling of the non-native corpora used for training the tonal models for pronunciation scoring. However, many learners' pronunciations often fall between two canonical tone categories, not belonging to a single category. As a consequence, the use of hard targets (one-hot target) in training deep models may reveal to be suboptimal. Thus, soft targets are here proposed for our acoustic model design. Soft targets were originally introduced in [30-32], who utilized the knowledge learned from large-size complex models to guide small-size DNN training by minimizing Kullback-Leibler (KL) divergence between the model's output and the soft targets. Unlike the above-mentioned work, which apply soft targets for model compression, we use soft targets to help resolve the hard-assignments of non-native tone labels. Compared with hard targets (one-hot targets), the posteriors in soft target are more suitable for describing non-native tone production, e.g., the pitch contour might be in between two canonical tone categories or do not resemble any canonical tonal categories. Finally, we also averaged tone scores produced by BLSTM models trained with hard and soft targets to seek the complementarity from modeling at the tone-target levels.

2. OVERVIEW OF THE DETECTION FRAMEWORK

Figure 2 shows the proposed mispronunciation detection system, which consists of training and testing stages. In training stage, we investigated two training paradigms: (1) conventional acoustic model training with hard targets; (2) proposed acoustic model training with soft targets. In testing stage, along with the trained acoustic model and tone-based ERN grammar, the input speech is decoded into lattices, where tone-related scores are fed into a GOP calculator to decide whether current tone is correctly pronounced.

2.1. Acoustic Model Training with Hard Targets

Deep acoustic models in speech applications are often trained with forced-aligned labels generated from available generative models, e.g. GMM-HMM [33], we refer to these target labels as hard targets, because the probability is concentrated in only a single senone (e.g., clustered, context-dependent sub-phonetic states as introduced in [33]). The cross entropy between hard targets and the outputs of deep acoustic models are often computed as follows:

$$L^{(CE)}(\theta) = -\sum_{t} \log y_{it} \tag{1}$$

where y_{it} is the ith output of the neural network at time t, and the index i refers to ground true class at time t.

2.2. Acoustic Model Training with Soft Targets

In contrast to hard targets (one-hot targets), soft targets are characterized by a set of senones, and their probabilities. Soft targest can be obtained using speech lattice - a compact encoding structure containing multiple recognition hypotheses, generated by an existing ASR system. Subsequently, KL divergence between soft targets and the outputs of the deep acoustic models is formulated as follows:

$$L^{(KL)}(\theta) = -\sum_{t}\sum_{i} \hat{y}_{it} \log \frac{\hat{y}_{it}}{y_{it}} \tag{2}$$

where $\hat{y}_{it}$ is the ith dimension value of soft target at time t. Gradient decent method is used to update parameter θ associated with deep models (e.g., DNN and BLSTM) to minimize abovementioned cross entropy and KL divergence.

2.3. Tone-based ERN Construction

Similar to phone-based ERN [34, 35] used for phone-level mispronunciation detection, tone-based ERN is constructed by expanding each toneless syllable into five different tonal syllables. Figure 3 shows an example of a tone-based ERN, where the toneless syllable sequence is "***zhong guo***". The tone-based ERN can be used as a grammar (language model) to constrain the search space, so that each node in the decoded speech lattice is only associated with the provided toneless syllable, yet with different tone marks. The acoustic scores stored in the edges between the abovementioned nodes indicate how likely the pronounced tonal syllable belongs to each tone category.

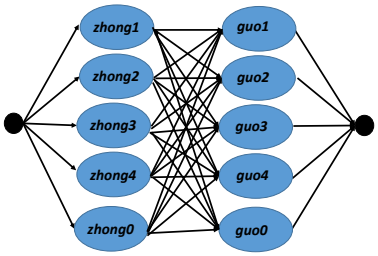

Figure 3: *An example of tone-based ERN*

2.4. GOP Score Calculation

After decoding the test speech sequences into a lattice structure and extracting segmental posteriors stored in the above-mentioned edges, we use Eq. (3) to calculate GOP score and measure whether the current syllable's tone is correctly pronounced.

$$GOP(p) = log \frac{P(p|o;t_s,t_e)}{max_{\{q\in Q\}}P(q|o;t_s,t_e)} \tag{3}$$

where o is the observed acoustic feature sequence starting from time t_s to t_e and p is the canonical tonal syllable, q is the competing tonal

syllable, and q is a set of possible units sharing the same toneless syllable as p, but owning different tone marks. A threshold is needed to verify whether the current tone is correctly pronounced.

3. EXPERIMENTS

3.1. Speech Corpora

Two speech corpora, (i) a native speech corpus from the Chinese National Hi-Tech Project 863 for Mandarin LVCSR system development [36], and (ii) a subset of a non-native speech corpus called iCALL [37], are mixed together to train phone-dependent tonal acoustic models. Our non-native test set contains 1662 utterances spoken by 30 leaners whose native languages contain English, French, Spanish, Italian, and Russian. There was no speaker overlaps between the training and test sets.

3.2. Evaluation Metrics

In this study, we adopt two common metrics, namely false acceptance rate (FAR) and false rejection rate (FRR), to evaluate mispronunciation detection performance. We do not employ the diagnostic accuracy as a third metric, because many mispronounced tones cannot be assigned to any of the single canonical tone categories. Therefore, traditional feedback related to tone-level substitution is not investigated.

$$FAR = \frac{FA}{N_H} * 100\% \quad (4)$$

$$FRR = \frac{FR}{N_C} * 100\% \quad (5)$$

where FA is the number of incorrect segments that are accepted by the system as correct, N_H is the total number of mispronunciations labeled by a human expert. FR is the number of correct tones that are misclassified as mispronounced ones, N_C is the total number of correct tone labeled by annotators.

3.3. Acoustic Modeling Setup

3.3.1. Cross-Entropy Based Hard Target Training Setup

The feature vector contains 23 dimensional FBANK coefficients, F0, the probability of voicing (POV) [38], and their derived velocity, and acceleration values. The acoustic model is trained with the open source Kaldi toolkit [39]: a CD-GMM-HMM acoustic model is initially trained with Maximum Likelihood (ML) criterion. Then the CD-DNN-HMM and CD-BLSTM-HMM models are trained by alignments provided by the CD-GMM-HMM system. The DNN model has six layers each containing 2048 sigmoid units. The input of DNN spans a window of 21 speech frames. This configuration is the same as our best system reported in work [8]. The BLSTM model has two hidden layers, 320 memory cells for each layer.

3.3.2. KL Divergence Based Soft Target Training Setup

Each utterance in the training set is first expanded into a tone-based ERN. Next, the above-trained CD-BLSTM-HMM and the tone-based ERN are used to decode utterances into lattices, where each frame is annotated with senone labels (which share the same toneless phone, but owning different tone marks) and their probabilities. We refer to the decoded lattice as soft target, which is subsequently used to retrain the CD-DNN-HMM and CD-BLSTM-HMM models.

3.4. Experimental Results and Discussions

3.4.1. Hard Target Experiments

After the GOP score computation, a tone-independent threshold is applied to find the equal error rate (EER), where FAR is equal to FRR. Figure 4 shows detection curves and EER points for both the baseline DNN system (blue curve) [8] and the proposed BLSTM system (orange curve).

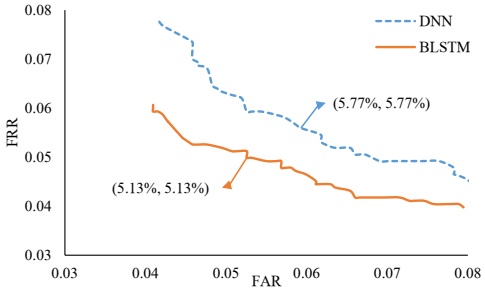

Figure 4: *DET curves for DNN and BLSTM tone mispronunciation detection systems with hard targets.*

Form Figure 4, we can see that the proposed BLSTM system consistently outperforms the DNN one, and the EER is reduced from 5.77% down to 5.13%. Long-range dependencies of acoustic and prosodic features captured by the BLSTM are therefore useful in tone verification, as expected. Figure 5 is an example of a testing utterance, where canonical transcription and human annotation are shown in the upper part. Its corresponding tone-related posteriors produced by DNN (denoted in blue color) and BLSTM (denoted in orange color) are summarized in Table 1, where each posterior characterizes how likely the non-native pronounced tone belongs to each tone category, e.g., the posteriors in lower right corner of Table 1 indicate that both DNN and BLSTM recognized that the non-native pronunciation of canonical *tone4* belongs to *tone4* class with 1.0 posterior probability (confidence score). Regarding to the non-native pronunciation of canonical *tone3*, although the human annotator and the BLSTM system agreed that it is correctly pronounced as *tone3*, the DNN-based systems gives a higher preference to *tone2* assigning a confidence score equal to 0.66 (as shown in the second column in Table 1). We think that *tone2* is preferred by the DNN-based system because its contour is highly similar *tone3* at the end of the syllables, e.g., the rising slope in Figure 1. The fixed-context input window of the DNN-based system therefore fails to capture the discriminative information existing at the beginning of the *tone3* syllable. That is, the duration of the pitch contour in *tone3* is around 0.7 second, which cannot be appropriately modeled by a DNN, even with long input window splices.

The long-range dependencies captured by BLSTM is also helpful in modeling tone-level coarticualtion influenced by 2, or 3 neighboring syllables. In Figure 6, a test utterance is shown, where the canonical transcription and human labeling are displayed in the upper part. We analyze the syllable where the canonical tone is tone1: Although the human labeling indicates it is mispronounced and BLSTM shows a 0.95 confidence score for *tone2* (suggesting it is mispronounced), the DNN-based system judges this tone as properly produced, assigning a 0.76 confidence score to it (see third column in Table 2). The DNN-based system wrongly gives a higher preference to *tone1* because the non-native pitch contour of the tone in question is similar to the canonical *tone1* (see Figure 1), where pitch trajectories are both straight lines, with no rising or falling. However, the pitch contour of the tone in question can also be a tone-

level co-articulation product of *tone2* or *tone3*, when the preceding tone is tone4 and the following tone is *tone1* [25, 40]. These results suggest that the context information captured by the fixed context window of a DNN cannot decide whether the pitch contour of the tone in question is canonical *tone1*'s realization or the product of tone-level co-articulation.

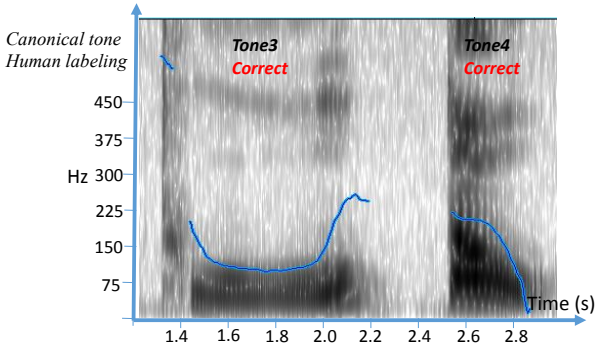

Figure 5: *Pitch contours and spectrograms of non-native pronunciation of canonical tone3 and tone4*

Table 1: *Tone-related posteriors produced by DNN (denoted in blue color) and BLSTM (denoted in orange color) for Figure 5.*

	Canonical Tone3	Canonical Tone4
Tone1	0.0/0.0	0.0/0.0
Tone2	0.66/0.01	0.0/0.0
Tone3	0.34/0.99	0.0/0.0
Tone4	0.0/0.0	1.0/1.0

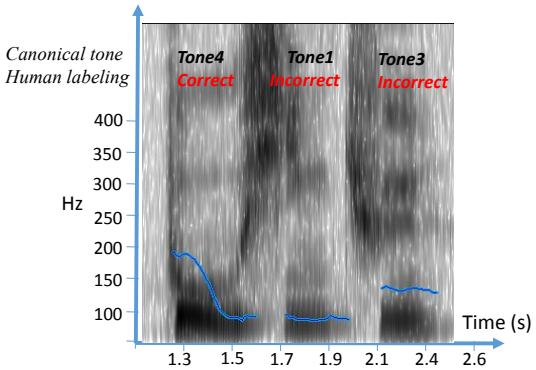

Figure 6: *Pitch contours and spectrograms of non-native pronunciation of canonical tone4, tone1 and tone3.*

Table 2: *Tone-related posteriors produced by DNN (denoted in blue color) and BLSTM (denoted in orange color) for Figure 6.*

	Canonical Tone4	Canonical Tone1	Canonical Tone3
Tone1	0.0/0.0	0.76/0.05	1.0/1.0
Tone2	0.0/0.0	0.13/0.95	0.0/0.0
Tone3	0.0/0.0	0.11/0.0	0.0/0.0
Tone4	1.0/1.0	0.0/0.0	0.0/0.0

3.4.2. Soft Target Experiments

Finally, the proposed BLSTM-based system consistently outperforms the baseline DNN-based system even when soft targets are used in training, as shown in Figure 7. Indeed, it should be pointed out that soft targets are beneficial for both DNN- and BLSTM-based systems. This observation indicates that soft targets could be more suitable for labeling irregular non-native tone pronunciations, namely L2 learners' tone realizations often fall between two canonical tone categories.

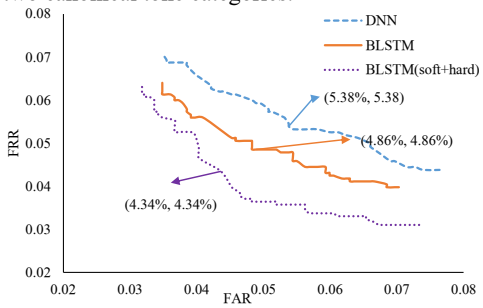

Figure 7: *Comparing tone mispronunciation detection performance of DNN-, BLSTM-based and combined systems.*

3.4.3. Posterior Score Combination Experiments

Although acoustic models trained with soft targets enhanced tone-level mispronunciation detection performance, there is still room for further improvement. Specifically, we averaged tone scores produced by BLSTM models trained with hard and soft targets to seek the complementarity from modeling at the tone-target levels. In doing so, we found that the EER of combined system is 4.34% as shown in Figure 7, about 24.78% relative error reduction was achieved, compared with our DNN baseline in Figure 4.

4. CONCLUSION AND FUTURE WORK

Through a series of systematic experiments, we have shown that performance of CAPT system for Mandarin tone can be significantly improved by properly choosing the neural architecture and the training scheme to best fit inherent features/peculiarities of the language at hand, namely (i) effect of long-term acoustic and prosodic dependencies in tonal languages, and (ii) effect of imprecise tonal pronunciation of L2 learners. We first have replaced DNN with BLSTM in tone-based extended recognition network (ERN) system [8, 9], which led to a significant performance improvement because of the better capability of a recurrent net to leverage information across wider temporal spans than a DNN architecture. Next, by observing that L2 learners' tone pronunciations are often between two canonical tone categories, we have adopted soft targets in modeling training, which led to a further enhancement of the recognition accuracy. The proposed framework significantly reduced the EER from 5.77% (hard-target, and DNN) to 4.86% (soft-target, and BLSTM). Finally, system combination produced a 24.78% relative reduction in the EER.

In this work, we have performed tone mispronunciation detection on sequences with no phonetic mispronunciations to comply with previous studies [8, 17]. For future work, we plan to relax this constraint and apply our approach to more challenging scenarios, where the spoken sequence has phonetic errors.

5. ACKNOWLEDGMENT

The first author was partially supported by a grant from the China Scholarship Council. The third author is partially supported by the NFR AULUS project. We would like to thank Beijing Language and Culture University SAIT Lab for helping us refining our non-native tone label in iCALL corpus.

6. REFERENCES

[1] W. Li, *et al*. "Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling." *In Proc. ICASSP, 2016.*

[2] W. Hu, *et al*. "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers." *Speech Communication*, 2015

[3] K. Li, X/ Qian, and H. Meng. "Mispronunciation detection and diagnosis in L2 english speech using multidistribution deep neural networks." *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2017.

[4] R. Tong, *et al*. "Context aware mispronunciation detection for Mandarin pronunciation training." *In Proc Interpeech, 2016*.

[5] H. Huang, *et al*. "Maximum F1-score discriminative training criterion for automatic mispronunciation detection." *IEEE/ACM Trans. on Audio, Speech and Language Processing, 2015.*.

[6] K. Imoto, *et al*. "Modeling and automatic detection of English sentence stress for computer-assisted English prosody learning system." *Seventh International Conference on Spoken Language Processing*. 2002.

[7] K. Li, *et al*. "Lexical stress detection for L2 English speech using deep belief networks." *In Proc. Interpeech, 2013*.

[8] W. Li, *et al*. "Using tone-based extended recognition network to detect non-native Mandarin tone mispronunciations." *In Proc. APSIPA, 2016.*

[9] J. Lin, *et al*. "Improving Mandarin Tone Recognition Based on DNN by Combining Acoustic and Articulatory Features Using Extended Recognition Networks." Journal of Signal Processing Systems, 2018.

[10]L. Zhang, *et al*. "Automatic detection of tone mispronunciation in Mandarin," *Chinese Spoken Language Processing*. Springer, 2006.

[11] J. Cheng, "Automatic tone assessment of non-native Mandarin speakers," *In Proc. Interpeech, 2012.*

[12] Y. B. Zhang, *et al*. "Detecting tone errors in continuous mandarin speech," *In Proc. ICASSP, 2008.*

[13] S. Wei, *et al*. "CDF-matching for automatic tone error detection in Mandarin CALL system," *In Proc. ICASSP, 2007.*

[14] J. Lin, *et al*. "Automatic pronunciation evaluation of non-native mandarin tone by using multi-level confidence measures," *In Proc. Interpeech, 2016*.

[15] Y. H. Guan, *et a*l. "Decision Tree Based Tone Modeling with Corrective Feedbacks for Automatic Mandarin Tone Assessment*," In Proc. Interpeech, 2010*.

[16] R. Tong, *et al*. "Tokenizing fundamental frequency variation for mandarin tone error detection." *In Proc. ICASSP, 2015.*

[17] R. Tong, N. F. Chen, B. Ma, and H. Li, "Goodness of Tone (GOT) for Non-native Mandarin Tone Recognition," *In Proc. Interpeech, 2015.*

[18] W. Hu, Y.Qian, and F. K. Soong. "A DNN-based acoustic modeling of tonal language and its application to Mandarin pronunciation training." *In Proc. ICASSP, 2014*.

[19] A. Ito, *et al.* "Automatic detection of English mispronunciation using speaker adaptation and automatic assessment of English intonation and rhythm." *Educational technology research, 2006.*

[20] K. Li, X. Wu, and H. Meng. "Intonation classification for L2 English speech using multi-distribution deep neural networks." *Computer Speech & Language*, 2017.

[21] N. Ryant, M. Slaney, M. Liberman, E. Shriberg, and J. Yuan, "Highly Accurate Mandarin Tone Classification in The Absence of Pitch Information," *in Proc. Speech Prosody, 2014*.

[22] N. Ryant, J. Yuan, and M. Liberman, "Mandarin tone classification without pitch tracking," *In Proc. ICASSP, 2014*.

[23] C. Cao, *et al*. "The preliminary study of influence on tone perception from segments." *In Proc. ISCSLP, 2016*.

[24] C. Cao, *et al*. "The Influence on Realization and Perception of Lexical Tones from Affricate's Aspiration." *In Proc. Interpeech, 2017*.

[25] Y. Xu, "Production and perception of coarticulated tones." The Journal of the Acoustical Society of America, 1994.

[26] R. Tsai, "Teaching and learning the tones of Mandarin Chinese." *Scottish Languages Review*, 2011.

[27] Y.-C. Hao, "Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers." *Journal of Phonetics*, 2012.

[28] Y. Wang, A. Jongman, and J. A. Sereno. "Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training." *The Journal of the Acoustical Society of America*, 2003.

[29] S. Hochreiter, and J. Schmidhuber. "Long short-term memory." Neural computation, 1997.

[30] J. Li, *et al*. "Learning small-size DNN with output-distribution-based criteria." *In Proc. Interpeech, 2014*.

[31] G. Hinton, O. Vinyals, and J. Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:2015.

[32] W. Chan, N. R. Ke, and I. Lane. "Transferring knowledge from a RNN to a DNN." *arXiv preprint arXiv:2015.*.

[33] G. E. Dahl, *et al*. "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition." *IEEE/ACM Transactions on audio, speech, and language processing, 2012*.

[34] H. Meng, *et al*. "Deriving salient learners' mispronunciations from cross-language phonological comparisons*," In Proc. ASRU, 2007*.

[35] W. K. Lo, S. Zhang, and H. M. Meng, "Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system*," In Proc. Interpeech, 2010*.

[36] S. Gao, *et al*, "Update of progress of sinohear: advanced Mandarin LVCSR system at NLPR," *In Proc. ICSLP, 2000*.

[37] 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.

[38] P. Ghahremani, *et al*. "A pitch extraction algorithm tuned for automatic speech recognition," *In Proc. ICASSP, 2014.*

[39] 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.*

[40] S. Duanmu, *The Phonology of Standard Chinese*, Oxford University Press, 2007.