A DNN-BASED ACOUSTIC MODELING OF TONAL LANGUAGE AND ITS APPLICATION TO MANDARIN PRONUNCIATION TRAINING

Wenping Hu[1,2,*] *Yao Qian*[2] *Frank K. Soong*[2]

[1]University of Science and Technology of China, Hefei, China
[2]Microsoft Research, Beijing, China
{v-wenh,yaoqian,frankkps}@microsoft.com

ABSTRACT

In this paper we investigate a Deep Neural Network (DNN) based approach to acoustic modeling of tonal language and assess its speech recognition performance with different features and modeling techniques. Mandarin Chinese, the most widely spoken tonal language, is chosen for testing the tone related ASR performance. Furthermore, the DNN-trained, tone-sensitive model is evaluated in automatic detection of mispronunciation among L2 Mandarin learners. The best DNN-HMM acoustic model of tonal syllable (initial and tonal final), trained with embedded F0 features, has shown improved ASR performance, when compared with the baseline DNN system of 39 MFCC features. The proposed system achieves better ASR performance than the baseline system, i.e., by 32% and 35% in relative tone error rate reduction and 20% and 23% in relative tonal syllable error rate reduction, for female and male speakers, respectively. In a speech database of L2 Mandarin learners (native speakers of European languages), 2% equal error rate reduction, from 27.5% to 25.5%, has been obtained with our DNN-HMM system in detecting mispronunciations, compared with the baseline system.

Index Terms— Computer-Aided Pronunciation Training, F0, Acoustic Model, Mandarin, Deep Neural Network

1. INTRODUCTION

In a tonal language, tone plays an important lexical role in addition to its relevance to speech prosody, i.e., words of the same syllables but different tones are lexically different. Therefore, acoustic modeling of a tonal language inevitably needs to utilize the tone relevant information for good speech recognition performance or for a Computer Assisted Language Learning (CALL) system. Among all tonal languages, Mandarin Chinese, the official language in China, is the mostly widely used tonal language in terms of speaking population. Also, due to the rapid economic development in China recently, Mandarin has become increasingly popular as a second language among speakers of different mother tongues. It is recently estimated that more than 40 million people around the world are studying Chinese [1]. Naturally, this phenomenal surge of Chinese learning creates a shortage of qualified teachers and language learning aids. CALL systems, powered by the advancement of speech technology, can bridge the gap between the supply and demand of Chinese language teachers and have become ubiquitous learning tools with handy smart phones, tablets, laptop computers, etc. The indispensable, lexical role played by the tonal patterns in Mandarin turns out to be a difficult hurdle for a foreign speaker whose mother tongue is non-tonal to perceive and to mimic the tone patterns. A high quality Computer Aided Pronunciation Training (CAPT) system is therefore useful for helping such L2 language learners.

Recently, a new machine learning algorithm called Deep Neural Network (DNN) has demonstrated significant speaker independent, continuous speech recognition performance improvement, compared with the conventionally trained GMM recognizers [2][3]. Lei [4] has incorporated tone-related posteriors, extracted from a multi-layer perceptron (MLP), with spectral features for conversational Mandarin speech recognition and obtained 2.5% character error rate reduction. Qian [5] has also applied the DNN to mispronunciation detection and diagnoses in L2 English language learners and obtained significant performance improvement. Extending this approach to pronunciation quality scoring has achieved similar performance improvement [6]. The DNN-based structure is efficient to decompose the input features into effective basis functions which in turn can be further discriminatively trained by a "soft-max" top layer of DNN to simulate the class posterior probabilities, e.g., the sub-phonemic "senones" units. The Goodness of Pronunciation (GOP) scores estimated from the DNN output correlate better with human expert's evaluations than the conventional GOP scores obtained with a conventional GMM-based system [6]. The frame posterior which can be computed directly without going through a forward-backward decoding lattice,is also advantageous for fast, on-line, multi-channel applications.

In this study, we investigate different ways to employ F0,

*Intern in Speech Group, Microsoft Research Asia

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

the main acoustic cue of tonal information, in DNN-based acoustic modeling of Mandarin speech. Its improved performance is assessed with both speech recognition and mispronunciation detection.

2. DNN-BASED TONE MODELING

Each Chinese character, which is a morpheme in written Chinese, is pronounced as a tonal syllable, i.e., a base syllable plus a lexical tone. All Mandarin syllables have a structure of (consonant)-vowel-(consonant), where only the vowel nucleus is an obligatory component. A mandarin syllable without tone, is referred as a base syllable here. By the convention of Chinese phonology, each base syllable can be divided into two parts: initial and final. The initial (onset) includes what precedes the vowel while the final includes the vowel (nucleus) and what succeeds it (coda)[7]. Most Mandarin initials are unvoiced and the tones are carried primarily by the finals in tonal syllables.

While acoustic features such as duration and energy also contribute partially to tone perception, F0 contour is the most important acoustic cue of Mandarin tone. In this study we focus only on F0 and its modeling. In the last few decades, the main stream for acoustic modeling of speech is GMM-HMM framework, where a GMM is used to approximate the distributions of continuous acoustic features. Different from spectral features, tone or F0 contour, is only quasi-continuous and F0 disappears in unvoiced segments. Two different approaches have been proposed to get around this discontinuity problem in modeling the tone. A heuristic approach is to interpolate F0 in unvoiced regions. The interpolation can be generated from a smooth function such as a quadratic spline [8], an exponential decaying function towards the running F0 average [9], or a probability density function (pdf) with a very large variance [10]. The other approach is to utilize a more advanced, mixed distribution to model the observation probability, e.g., multi-space distribution (MSD), proposed by Tokuda[11] to model the discontinuous pitch contours statistically. It has been successfully applied to HMM-based speech synthesis [11] and recognition [7].

In this study, we employ a DNN-based framework for embedded tone modeling, where tone features are appended to the spectral features, for modeling Mandarin speech. A Deep Neural Network (DNN) is a feed-forward, artificial neural network with multiple hidden layers between its input and output. For each hidden unit j, a function, typically a logistic one, is used to map all input from the lower layer, x_j, to a scalar state, y_j, which is then fed to the upper layer.

$$y_j = logistic(x_j) = \frac{1}{1+e^{-x_j}}; \quad x_j = b_j + \sum_i y_i w_{ij} \quad (1)$$

where b_j is the bias of unit j; i, the unit index of lower layer; w_{ij}, the weight on the connection between unit j and unit i in the layer below. For multi-class classification, a "soft-max" nonlinear function is used to converts the inputs, x_j, into a class probability, p_j, given in eq (2), where k is an index over all classes.

$$p_j = \frac{exp(x_j)}{\sum_k exp(x_k)} \quad (2)$$

All weights and bias are initialized in DBN pre-training [12], and then discriminatively trained by optimizing the cross entropy between the target probability and actual output of "soft-max" with the Back-Propagation (BP) procedure [13].

There are many advantages to embed the F0 contour in DNN over GMM. First, since tone is a suprasegmental feature which goes far beyond the time span of a single frame, a longer window of observations is necessary for capturing the supersegmental characteristics of tone, hence the resultant modeling accuracy. The DNN structure is more conveniently set for augmenting a longer time span of adjacent frames than the GMM in modeling. Further, there is no underlying assumption of the distribution and modality for input data in the DNN, e.g., continuous and binary features can be augmented and modeled together naturally, while it is impossible or cumbersome to deal with those heterogeneous features in GMM. For example, complex modeling methods like interpolation or MSD have to be adopted. In this study, we use only extracted F0 values in its logarithmic form (F0 in unvoiced segments is set as zero, with no interpolation) together with the spectral features for the DNN based acoustic modeling.

3. EXPERIMENTS AND RESULTS

In our experiments, the DNN-based tone embedded models are trained and evaluated on tone related recognition performance and then tested on their ability in assessing a learner's pronunciation quality.

3.1. Tonal Syllable Recognition

A speaker-independent, large vocabulary and continuous Mandarin speech database (BJ2003) [7] is used for acoustic modeling. In total, there are 490 speakers (gender balanced) and each speaker was requested to read through a set of Chinese text, including modern novels and classical Chinese writings. The training data contains about 66 hrs of speech recordings from 230 male and 230 female speakers. 4,000 utterances recorded by the remaining 16 male and 14 female speakers are used for testing.

All speech signals are sampled in 16k Hz. The MFCC features, extracted with a 25ms hamming window, shifted every 10ms, consist of 13 MFCCs. The extraction of F0 is done on a short-time basis by applying the robust algorithm for pitch tracking (RAPT) [14]. MFCC, F0 and their first and second-ordered time differences are concatenated together and used as input features for model training.

Table 1. *Experimental Configurations*

System	MFCC	F0	Δ F0	$\Delta\Delta$F0	U/V
DNN-A	$\surd$				
DNN-B	$\surd$	interp.	$\surd$	$\surd$	
DNN-C	$\surd$	raw	$\surd$	$\surd$	
DNN-D	$\surd$		$\surd$	$\surd$	$\surd$
DNN-E	$\surd$		$\surd$	$\surd$	

To investigate how to model F0 in a DNN-based framework, we configure the experiments as shown in Table 1 with different features and preprocessing options. In the table,"$\surd$" denotes the corresponding feature is selected in a specific model experiment; U/V is a binary voicing flag; under F0, interpolation preprocessing can be selected or not. In DNN-B, we use an exponential decaying function [9] to interpolate pitch in unvoiced speech regions. In DNN-C, raw log F0 is used, i.e., static F0 and its deltas are set as zero in unvoiced regions, and logarithmic F0 is only used in voiced regions, whose deltas are calculated by duplicating and extending two frames outward at the U/V boundaries.

Gender-dependent DNN models are trained. All DNN models with the same structure: 3 hidden layers, 2k nodes for each hidden layer, and 3,008 output "senone" states, are trained with the same experimental configurations, i.e. same epochs, learning rates and other training parameters. The input of DNN is an augmented, 11 frame super-vector including 5 preceding frames, the current frame and 5 succeeding frames. The phone set used is Phn187 [10], consisting of 187 "phones", i.e., initials and tonal finals. Tonal final is a tone embedded unit, i.e., any model trained with Phn187 are trained with corresponding canonical tone labels, with or without F0 features. In recognition decoding, a free tonal-syllable loop grammar is used. The tonal syllable recognition performance between the conventional GMM and DNN based HMMs is shown in Table 2.

Table 2. *Tonal Syllable Error Rate for different models*

	Male (%)	Female (%)
MFCC-39	46.80	45.24
MSD-44-2S	37.98 (+18.9%)	34.77 (+23.1%)
DNN-A	35.40 (+24.4%)	32.88 (+27.3%)

Compared with a baseline system, the relative improvements are shown in parentheses correspondingly. Between the two previously trained GMM models, MFCC-39 and MSD-44-2S, MSD-44-2S yields a better recognition performance by modeling the F0 with MSD [7]. A new DNN trained model, DNN-A, which uses the same 39 MFCC spectral features as the MFCC-39 system, outperforms the two GMM models by a large margin, even without employing F0 features. DNN-A will be used as our baseline DNN model in later experiments.

In Tables 3, 4 and 5, we compare the tonal-syllable error rate (TSER), tone error rate (TER) and base-syllable error rate (BSER) of different DNN models. As shown in Table 3, when F0 is embedded into DNN modeling (systems DNN-B and C), we can significantly improve the tonal syllable recognition performance over the baseline DNN-A where no F0 is used. Interpolating F0 (DNN-B) or not (DNN-C) doesn't seem to matter that much in recognition performance and their relative improvement of TSER over the baseline DNN-A is slightly perturbed around 23% and 20% for male and female speakers, respectively. In Table 4, the corresponding relative TER improvements are 35% and 32% for male and female testing sets. Table 5 shows that by integrating F0 features into DNN we can also improve the recognition performance of base-syllable error rate (BSER), though by a smaller margin. Comparing with DNN-B, DNN-C yields a slightly better performance in base syllable recognition while keeping almost the same performance on tone and tonal syllable recognition. We conjecture that raw F0 can be better utilized by DNN-C training for discriminating voiced against unvoiced segments, while the artificial F0 interpolation in DNN-B, could have introduced some adverse, though minor, effect to spectrum modeling, hence degrades slightly, its error rate in base-syllable recognition. Since the relative dynamic range of logarithmic F0 for voiced segments is small, most of the static F0 points in DNN-C are distributed in a small region. In DNN-based modeling, all input features are normalized to a zero mean and unity variance distribution. We think the static F0 value is mainly functioned as a binary, U/V decision flag and its numerical value has little effect in tone discrimination. To further check this conjecture on the roles played by static F0 and their deltas in tone recognition, we use DNN-D and DNN-E. In DNN-D we replace the static F0 with a binary U/V flag and in DNN-E only use two F0

Table 3. *TSER for different tone embedded DNN models*

	Male (%)	Female (%)
DNN-A	35.40	32.88
DNN-B	27.36 (+22.7%)	26.29 (+20.0%)
DNN-C	27.21 (+23.1%)	26.36 (+19.8%)

Table 4. *TER for different tone embedded DNN models*

	Male (%)	Female (%)
DNN-A	25.69	22.92
DNN-B	16.65 (+35.2%)	15.54 (+32.2%)
DNN-C	16.92 (+34.1%)	15.64 (+31.8%)

Table 5. *BSER for different tone embedded DNN models*

	Male (%)	Female (%)
DNN-A	19.93	18.09
DNN-B	18.14 (+9.0%)	17.07 (+5.6%)
DNN-C	17.50 (+12.2%)	16.69 (+7.7%)

deltas without the static F0. Their TER results are shown in Table 6. It is observed that without the static F0 in DNN, the TER recognition performance reduction is negligible. and dynamic F0 features tend to be more important than static F0 in recognizing Mandarin tones. It reconfirms that tones are recognized more by their dynamic shapes characterized by F0 deltas than their static values.

Table 6. *TER of DNN-C, D and E*

	Male (%)	Female (%)
DNN-C	16.92	15.64
DNN-D	17.15 (−1.4%)	15.72 (−0.5%)
DNN-E	17.43 (−3.0%)	15.98 (−2.2%)

3.2. Tone Pronunciation Evaluation

A L2 Mandarin language learning corpus, recorded by 295 European speakers whose mother tongues are mainly English or French, is used to evaluate the performance of thus built Mandarin CALL system. Each speaker reads about 300 utterances of isolated words, continuous phrases and sentences. A randomly chosen subset of 1,000 utterances (gender balanced) is rated by two trained phoneticians at initial/tonal final level. In total, there are 2,102 mispronounced initial/tonal finals among all 8,886 units.

Since utterances are recorded by people from different countries with different mother tongues, an L1 independent approach is adopted. The DNN based Log Likelihood Ratio between correct and competing models [6], is used as Goodness of Pronunciation (GOP) measure to evaluate the pronunciation quality.The GOP score of a phone p is defined in eq (3).

$$GOP(p) = PLTF\{\frac{1}{t_e - t_s} \cdot \{\sum_{t=t_s}^{t_e} logP(o_t|p) - \max_{\{q \in Q, q \neq p\}} \sum_{t=t_s}^{t_e} logP(o_t|q)\}\} \quad (3)$$

where $PLTF(\cdot)$ is a piecewise linear function to convert the ratio score to a percentage grading; o_t is the argument input observations of the frame t; t_s and t_e are the start and end frame indices of phone p, respectively; $P(o_t|p)$ is the likelihood approximately calculated as the division between the output of our DNN model $P(s_t^p|o_t)$ and the prior of label s_t^p, where s_t^p is the "senone" label of frame t generated by force alignment with the given phone p; Q is the set of all initials/tonal finals for the canonic initial/tonal final p. Our paper [6] gives more detail description about the GOP estimation algorithms.

We compare the detection performance of two acoustic models, DNN-A and DNN-C. The GOP score for each initial/tonal final is calculated as eq (3), while their corresponding human label, correct pronunciation or mispronunciation, can be obtained by comparing its canonical transcription with spoken transcription which is labeled by phoneticians. A binary decision of correct pronunciation or mispronunciation is made for each initial or tonal final with a pre-defined threshold. By changing the threshold, we obtain the Receiver Operating Characteristic (ROC) Curve for the two systems as depicted in Fig.1. The performance of DNN-C is consistently better than DNN-A. The Equal Error Rate (EER) is improved by 2.0% from 27.5% down to 25.5% with the embedded F0 features. It is slightly disappointing that by incorporating the most relevant tone feature F0 in DNN, we managed to improve the detection EER by only 2%. However, we should be aware that DNN is a very powerful machine learning algorithm which can apparently infer some tone information only from the 39 MFCC spectral features. The recognition performance in tonal syllable error rate comparison (shown in Table 2) between MFCC-39 and DNN-A systems, where both systems use the same 39 MFCC features, has also confirmed the superiority of the DNN's power of this inference. In the future, more acoustic features, e.g., energy, duration, will be investigated.

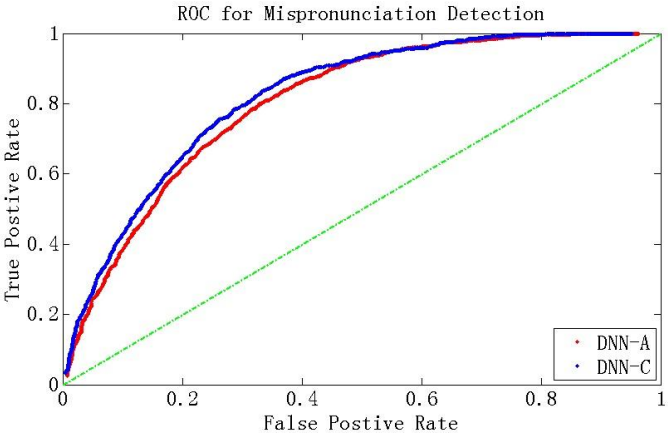

Fig. 1. Initial and tonal final mispronunciation detection

4. CONCLUSION

We investigate a DNN-based approach to acoustic modeling of tonal language. Different tone-embedded modeling techniques are compared and their performances in tone related Mandarin speech recognition are evaluated. DNN-HMM is also assessed in its ability on detecting L2 Mandarin learners' mispronunciations. The best tone embedded DNN-HMM can improve relative tone recognition error rates by 32% and 35%, or relative tonal syllable error rates by 20% and 23%, for female and male speakers, respectively, compared with the corresponding DNN-HMM without embedding F0. For L2 Mandarin learner's speech data, an improvement of 2% EER of mispronunciation detection reduction, from 27.5% to 25.5%, can be obtained, by comparing two DNN-based systems without/with embedded tone modeling.

5. REFERENCES

[1] "40 million people worldwide study Chinese," http://english.people.com.cn/90001/90782/90872/7112508.html.

[2] George E. Dahl, Dong Yu, Li Deng and Alex Acero, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, pp. 30–42, 2012.

[3] Frank Seide, Gang Li, Xie Chen and Dong Yu, "Feature engineering in Context-Dependent Deep Neural Networks for conversational speech transcription," in *ASRU*, 2011, pp. 24–29.

[4] Xin Lei, Mei-Yuh Hwang and Mari Ostendorf, "Incorporating tone-related MLP posteriors in the feature representation for Mandarin ASR," in *INTERSPEECH*, 2005, pp. 2981–2984.

[5] Xiaojun Qian, Helen Meng and Frank K. Soong, "The Use of DBN-HMMs for Mispronunciation Detection and Diagnosis in L2 English to Support Computer-Aided Pronunciation Training," in *INTERSPEECH*, 2012.

[6] Wenping Hu, Yao Qian and Frank K. Soong, "A New DNN-based High Quality Pronunciation Evaluation for Computer-Aided Language Learning (CALL)," in *INTERSPEECH*, 2013, pp. 1886–1890.

[7] Yao Qian and Frank K. Soong, "A Multi-Space Distribution (MSD) and two-stream tone modeling approach to Mandarin speech recognition," *Speech Communication*, vol. 51, pp. 1169–1179, 2009.

[8] Daniel Hirst and Robert Espesser, "Automatic Modelling Of Fundamental Frequency Using A Quadratic Spline Function," *Travaux de lInstitut de Phontique dAix*, vol. 15, pp. 75–85, 1993.

[9] C. Julian Chen, Ramesh A. Gopinath, Michael D. Monkowski, Michael A. Picheny and Katherine Shen, "New methods in continuous Mandarin speech recognition," in *EUROSPEECH*, 1997, pp. 1543–1546.

[10] Eric Chang, Jianlai Zhou, Shuo Di, Chao Huang and Kai-Fu Lee, "Large Vocabulary Mandarin Speech Recognition with Different Approaches in Modeling Tones," in *INTERSPEECH*, 2000, pp. 983–986.

[11] Keiichi Tokuda, Takashi Mausko, Noboru Miyazaki and Takao Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. & Syst.*, vol. E85-D, pp. 455–464, 2002.

[12] Geoffrey E. Hinton, Simon Osindero and Yee-Whye Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[13] David E. Rumelhart, Geoffrey E. Hinton and Ronald J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[14] David Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, pp. 495–518, 1995.