LETTER

Discriminative pronunciation modeling using the MPE criterion

Meixu SONG[†a)], **Jielin PAN**[†], ***Nonmembers***, **Qingwei ZHAO**[†], ***Member***, ***and*** **Yonghong YAN**[†], ***Nonmember***

SUMMARY Introducing pronunciation models into decoding has been proven to be benefit to LVCSR. In this paper, a discriminative pronunciation modeling method is presented, within the framework of the Minimum Phone Error (MPE) training for HMM/GMM. In order to bring the pronunciation models into the MPE training, the auxiliary function is rewritten at word level and decomposes into two parts. One is for co-training the acoustic models, and the other is for discriminatively training the pronunciation models. On Mandarin conversational telephone speech recognition task, compared to the baseline using a canonical lexicon, the discriminative pronunciation models reduced the absolute Character Error Rate (CER) by 0.7% on LDC test set, and with the acoustic model co-training, 0.8% additional CER decrease had been achieved.

key words: *automatic speech recognition, pronunciation models, discriminative training, Mandarin conversational speech recognition*

1. Introduction

Current LVCSR technology aims at transferring real-world speech to sentence. Due to data sparsity, it is almost impossible to find a sufficiently direct conversion between speech and sentence. Therefore, this conversion is divided into three parts, as show in Fig. 1: (a) the conversion between speech feature vectors and subwords (phones for example) described by Acoustic Models (AMs); (b) the conversion between words and sentence described by Language Model (LM); and (c) the conversion between subwords and words described by a lexicon. We consider that a lexicon is composed of three parts: words, pronunciations, and Pronunciation Models (PMs). PMs contain the pronunciation probabilities of each word in the lexicon. In many LVCSR systems, the lexicon is hand-crafted, that usually means the pronunciations are in canonical forms, and the probability in PMs could be considered as constant 1. To automatically learn a lexicon, earlier studies have explored the data-driven pronunciation learning and PM training methods.

As to pronunciation learning, earlier work [1] presented a discriminative pronunciation learning method using phonetic decoder and minimum classification error criterion. And previous work [2], [3] made use of a state-of-the-art letter-to-sound (L2S) system based on joint-sequence modeling [4] to generate pronunciations. Specifically for Mandarin pronunciation learning, the pronunciation variants of each constituent character in a word were enumerated to construct a pronunciation dictionary in [5]. This method is used to generate pronunciations for words in this paper, and the implementation details will be described.

Fig. 1 The conversions between speech and sentence

As to PM training, in [2], [3], a pronunciation mixture model (PMM) was presented by treating pronunciations of a particular word as components in a mixture, and the distribution probabilities were learned by maximizing the likelihood of acoustic training data. By contrast, in our work we modify the auxiliary function of the standard MPE training [6] to incorporate PMs. By doing so, a discriminative pronunciation modeling method using minimum phone error criterion is proposed, called MPEPM. In this method, the acoustic models and pronunciation models are co-trained in an iterative fashion under the MPE training framework.

In the experiment on two Mandarin conversational telephone speech test sets, compared to the baseline using a canonical lexicon, the proposed method has 1.5% and 1.1% absolute decrease in CER respectively.

The rest of the paper is organized as follows: Section 2 gives a brief introduction of PMs in speech recognition. In Section 3, a detailed derivation of incorporation of PMs into MPE training is presented. Section 4 reports experimental results, and Section 5 gives a conclusion of this paper.

2. Pronunciation Models

With PMs considered in speech recognition, the most likely words sequence using Viterbi approximation is [1], [7]:

$$\begin{aligned}\widehat{W} &= \arg\max_{W} P(O|W)P(W) \\ &\cong \arg\max_{W,B} P(O|B)P(B|W)P(W) \end{aligned} \quad (1)$$

where O is the sequence of acoustic observations, W is a sequence of hypothesized words, and B is the sequence of possible pronunciations corresponding W. $P(B|W)$ is calculated by PMs. Suppose that PMs is context independent, then $P(B|W)$ can be written as:

$$P(B|W) = P(b_1, \ldots, b_{k_r}|w_1^r, \ldots, w_{k_r}^r) = \prod_{j=1}^{k_r} P(b_j|w_j^r)$$

$P(b_j|w_j^r)$ is the probability that the j-th word in the r-th word

Manuscript received October 16, 2014.
Manuscript revised October 16, 2014.

[†]The authors are with the Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, China.
a) E-mail: songmeixu@hccl.ioa.ac.cn

DOI: 10.1587/transinf.E0.D.1

sequence with k_r words, is pronounced as b_j.

3. Incorporate PMs into MPE training

To train PMs from speech corpus, the possible pronunciations sequences for each training utterance are usually required. In [3], these are obtained by decoding the N-best list of pronunciations. By contrast, the possible pronunciations sequences are already contained in lattices used in the standard MPE training. Thus, the incorporation of PMs into MPE training is investigated. The MPE objective function is [6]:

$$\mathcal{F}_{\mathrm{MPE}} = \sum_{r=1}^{R} \frac{\sum_W P_\lambda(O_r|W)P(W)\mathrm{A}(W)}{\sum_W P_\lambda(O_r|W)P(W)} \tag{2}$$

where $\mathrm{A}(W)$ represents the phone accuracy calculation function. To make Eq. (2) tractable, the auxiliary function of the MPE objective function is:

$$\mathcal{H}_{\mathrm{MPE}}(\lambda,\lambda') = \sum_{r=1}^{R}\sum_{q=1}^{Q_r} \gamma_q^{\mathrm{MPE}}\Big|^{(\lambda=\lambda')} \log P(q) \tag{3}$$

with

$$\begin{aligned}\gamma_q^{\mathrm{MPE}} &= \frac{\partial \mathcal{F}_{\mathrm{MPE}}}{\partial \log P(q)} \\ &= \gamma_q(c^r(q) - c^r_{avg})\end{aligned} \tag{4}$$

where $P(q)$ is the likelihood of the speech data aligned to phone arc q, γ_q is the posterior probability of the phone arcs q in current lattice, $c^r(q)$ is the average phone accuracy of paths passing through the phone arcs q, and c^r_{avg} is the average phone accuracy of all paths in the lattice of the r-th training utterance [6].

To incorporate PMs, as in Eq. (1), $P(O|W)$ is expanded to $P(O|B)P(B|W)$, then the MPE objective function is:

$$\mathcal{F}_{\mathrm{MPE}} = \sum_{r=1}^{R} \frac{\sum_W \sum_B P_\lambda(O_r|B)P(B|W)P(W)\mathrm{A}(W)}{\sum_W P_\lambda(O_r|B)P(B|W)P(W)}$$

We rewrite auxiliary function (Eq. (3)) at word level and incorporate pronunciation probability $P(b|w)$ as:

$$\mathcal{H}_{\mathrm{MPE}}(\lambda,\lambda') = \sum_{r=1}^{R}\sum_{w=1}^{W_r^\#} \gamma_{(w,b)}^{\mathrm{MPE}}\Big|^{(\lambda=\lambda')} \log P(b)P(b|w) \tag{5}$$

with

$$\begin{aligned}\gamma_{(w,b)}^{\mathrm{MPE}} &= \frac{\partial \mathcal{F}_{\mathrm{MPE}}}{\partial \log P(b)P(b|w)} \\ &= \gamma_{(w,b)}(c^r(w) - c^r_{avg})\end{aligned}$$

where (w,b) represents word w pronunciated as b. $P(b)$ is the likelihood of the speech data aligned to word arc (w,b). $W_r^\#$ is the words set in the lattice of the r-th utterance. Accordingly, $r_{(w,b)}$ is the posterior probability of the word arc (w,b) in current lattice. $c^r(w)$ is the average phone accuracy of paths passing through the word arc (w,b), and c^r_{avg} is the average phone accuracy of all paths in lattice of the r-th training utterance. Eq. (5) is based on a sum over word arcs $w = 1 \ldots W_r^\#$, each with start and end times.

By expanding $\log P(b)P(b|w)$ to $\log P(b) + \log P(b|w)$, Eq. (5) decomposes into two parts: AM co-training and PM training. The analyses of these two parts are as follows.

3.1 Co-train AMs

Suppose the pronunciation b for word w consists of phones $q_1^w \ldots q_{n_w}^w$, then

$$\log P(b) = \sum_{i=1}^{n_w} \log P(q_i^w)$$

where $P(q_i^w)$ is the likelihood of the data aligned to phone arc q_i^w. If the duration of q_i^w and q in Eq. (3) is equal, then $P(q_i^w) = P(q)$.

Thus, Eq. (5) becomes:

$$\begin{aligned}\mathcal{H}_{\mathrm{MPE}}(\lambda,\lambda') = &\sum_{r=1}^{R}\sum_{w=1}^{W_r^\#} \gamma_{(w,b)}^{\mathrm{MPE}} \sum_{i=1}^{n_w} \log P(q_i^w) \\ &+ \sum_{r=1}^{R}\sum_{w=1}^{W_r^\#} \gamma_{(w,b)}^{\mathrm{MPE}} log P(b|w)\end{aligned} \tag{6}$$

As the paths passing through word arc (w,b) are equal to those passing through any phone arc in word arc (w,b), namely for any $q_i^w \in (w,b)$:

$$\begin{aligned}\gamma_{(w,b)} &= \gamma_{q_i^w} \\ c^r(w) &= c^r(q_i^w) \\ \gamma_{(w,b)}^{\mathrm{MPE}} &= \gamma_{q_i^w}(c^r(q_i^w) - c^r_{avg})\end{aligned} \tag{7}$$

Then the first part of Eq. (6) is:

$$\begin{aligned}&\sum_{r=1}^{R}\sum_{w=1}^{W_r^\#} \gamma_{(w,b)}^{\mathrm{MPE}} \sum_{i=1}^{n_w} \log P(q_i^w) \\ &\qquad = \sum_{r=1}^{R}\sum_{w=1}^{W_r^\#}\sum_{i=1}^{n_w} \gamma_{q_i^w}(c^r(q_i^w) - c^r_{avg}) \log P(q_i^w)\end{aligned} \tag{8}$$

To keep statistics calculation consistent with that in the standard MPE training, we will demonstrate the above formula (Eq. (8)) is equal to the original auxiliary function (Eq. (3)) with γ_q^{MPE} calculated as:

$$\begin{aligned}\gamma_q^{\mathrm{MPE}} &= \frac{\partial \mathcal{F}_{\mathrm{MPE}}}{\partial \log P(q)} \\ &= \frac{\sum_{W_i, q\in W_i} P_\lambda(O_r|B_i)P(B_i|W_i)P(W_i)\mathrm{A}(W_i)}{\sum_{W_j} P_\lambda(O_r|B_j)P(B_j|W_j)P(W_j)} \\ &- \frac{\sum_{W_i, q\in W_i} P_\lambda(O_r|B_i)P(B_i|W_i)P(W_i)}{\sum_{W_j} P_\lambda(O_r|B_j)P(B_j|W_j)P(W_j)} \\ &\qquad \cdot \frac{\sum_{W_i} P_\lambda(O_r|B_i)P(B_i|W_i)P(W_i)\mathrm{A}(W_i)}{\sum_{W_j} P_\lambda(O_r|B_j)P(B_j|W_j)P(W_j)}\end{aligned} \tag{9}$$

The first part of Eq. (9) is:

$$\begin{aligned}&\frac{\sum_{W_i,q\in W_i} P_\lambda(O_r|B_i)P(B_i|W_i)P(W_i)\mathrm{A}(W_i)}{\sum_{W_j} P_\lambda(O_r|B_j)P(B_j|W_j)P(W_j)}\\ &= \frac{\sum_{q_i^w\in q}\sum_{W_i,q_i^w\in W_i} P_\lambda(O_r|B_i)P(B_i|W_i)P(W_i)\mathrm{A}(W_i)}{\sum_{W_j} P_\lambda(O_r|B_j)P(B_j|W_j)P(W_j)}\\ &= \sum_{q_i^w\in q} c^r(q_i^w)\cdot\gamma_{q_i^w}\end{aligned} \tag{10}$$

The second part of Eq. (9) equals to:

$$\begin{aligned}&\frac{\sum_{q_i^w\in q}\sum_{W_i,q_i^w\in W_i} P_\lambda(O_r|B_i)P(B_i|W_i)P(W_i)}{\sum_{W_j} P_\lambda(O_r|B_j)P(B_j|W_j)P(W_j)}\\ &\cdot\frac{\sum_{W_i} P_\lambda(O_r|B_i)P(B_i|W_i)P(W_i)\mathrm{A}(W_i)}{\sum_{W_j} P_\lambda(O_r|B_j)P(B_j|W_j)P(W_j)}\\ &= \sum_{q_i^w\in q}\gamma_{q_i^w}\cdot c^r_{avg}\end{aligned} \tag{11}$$

From Eq. (9)(10)(11) we have:

$$\gamma_q^{\mathrm{MPE}} = \sum_{q_i^w\in q}\gamma_{q_i^w}\cdot(c^r(q_i^w) - c^r_{avg}) \tag{12}$$

Finally, from Eq. (3)(12), we know Eq. (3) is a sum of $\gamma_{q_i^w}(c^r(q_i^w) - c^r_{avg})\log P(q_i^w)$ over phone arcs, while Eq. (8) is a sum of the same thing over word arcs. The results are equal, namely:

$$\begin{aligned}&\sum_{r=1}^{R}\sum_{w=1}^{W_r^\#}\gamma_{(w,b)}^{\mathrm{MPE}}\sum_{i=1}^{n_w}\log P(q_i^w)\\ &= \sum_{r=1}^{R}\sum_{q=1}^{Q_r}\gamma_q^{\mathrm{MPE}}\log P(q)\\ &= \sum_{r=1}^{R}\sum_{q=1}^{Q_r}\sum_{q_i^w\in q}\gamma_{q_i^w}\cdot(c^r(q_i^w) - c^r_{avg})\log P(q)\end{aligned}$$

Therefore, using γ_q^{MPE} calculated by Eq. (9), AMs are co-trained with PMs without changing the MPE framework.

3.2 MPEPM

From Eq. (6)(7), we get the objective function of PMs using minimum phone error criterion:

$$\sum_{r=1}^{R}\sum_{w=1}^{W_r^\#}\gamma_{(w,b)}^{\mathrm{MPE}} log P(b|w) \tag{13}$$

with constraints:

$$\sum_b P(b|w) = 1 \tag{14}$$

$$P(b|w)\in(0,1] \tag{15}$$

We define

$$\gamma_{(w,b)}^{\mathrm{num}} = \gamma_{(w,b)}^{\mathrm{MPE}}\ if\ \gamma_{(w,b)}^{\mathrm{MPE}}\geq 0$$
$$\gamma_{(w,b)}^{\mathrm{den}} = \gamma_{(w,b)}^{\mathrm{MPE}}\ if\ \gamma_{(w,b)}^{\mathrm{MPE}} < 0$$

Referring to the auxiliary function used to update weight in the MPE training [6], we use an auxiliary function for Eq. (13), that is:

$$\sum_{r=1}^{R}\sum_{w=1}^{W_r^\#}\gamma_{(w,b)}^{\mathrm{num}}\log P(b|w) - \frac{\gamma_{(w,b)}^{\mathrm{den}}}{P'(b|w)}P(b|w) \tag{16}$$

By maximizing this auxiliary function, the objective function in Eq. (13) is optimised with constraints of Eq. (14)(15). The detailed proofs could be found in [6]. For all (w,b), set $P^{(0)}(b|w) = P'(b|w)$, where $P'(b|w)$ is the probability in the former PMs. And the iterative formula is as follows, in the (p+1)-th iteration:

$$P^{(p+1)}(b|w) = \frac{\gamma_{(w,b)}^{\mathrm{num}} + k_b P^{(p)}(b|w)}{\sum_b \gamma_{(w,b)}^{\mathrm{num}} + k_b P^{(p)}(b|w)}$$

with

$$k_b = \left(\max_b \frac{\gamma_{(w,b)}^{\mathrm{den}}}{P'(b|w)}\right) - \frac{\gamma_{(w,b)}^{\mathrm{den}}}{P'(b|w)}$$

The values of $P^{(p+1)}(b|w)$ after 100 iterations are used as the updated pronunciation probabilities.

The above two subsections have shown the incorporation of PMs into the MPE training. Through this, a discriminative pronunciation modeling method is presented.

4. Experiments and Results

4.1 Construct Pronunciation Dictionary

We utilized the method employed in [5] to construct a pronunciation dictionary for 43k Chinese words. A character pronunciation dictionary with 7.8k pronunciations for 6.7k Chinese characters was used, to construct a full pronunciations set with 85k pronunciations. After performing a forced alignment of the acoustic training data, a 0.5 threshold relative to the maximum frequency of pronunciations of every word was set to prune out low frequent pronunciations. Finally, the pronunciation dictionary used to train PMs consisted of 47k pronunciations. The frequencies of remaining pronunciations of every word are normalized to form the initial pronunciation models.

4.2 Baseline System

Experiments were carried out on Mandarin conversational speech recognition task. The acoustic training data is about 400 hours, consisted of two parts. One is from LDC database including CallHome&CallFriend (45.9 hours), and LDC04[†] (150 hours) training sets. The other part is 200

[†]LDC04 was collected by Hong Kong University of Science and Technology (HKUST) in 2004

Table 1 Results in CER (%).

	HTest04	GDTest
baseline	49.7	50.8
MPEPM	49.0	50.5
co-train AMs & MPEPM	48.2	49.7

hours speech data collected by ourself. All the data were recorded through the landline telephone with local service in the real world with environmental noise. All utterances are in Chinese Mandarin and in spontaneous style.

There were three steps for the front-end process, First, a reduced bandwidth analysis, 60-3400 Hz, was used to generate 56-dimensional feature vectors, which consist of 13-dimensional PLP and smoothed F0 appended with the first, second and third order derivatives. Next, utterance-based cepstra mean and variance normalization (CMS/CVN) was applied. Finally, a heteroscedastic linear discriminant analysis (HLDA) [8] was directly applied to projected 56-dimensional feature vectors into 42-dimensions.

The phone set for HMMs modeling consists of 179 tonal phones. The final HMMs are cross-word triphone models with 3-state left-to-right topology, which are trained via the Minimum Phone Error (MPE) criteria [6]. A robust state clustering with phonetic decision trees is used [9], and finally 7995 tied triphone states are empirically determined with 16-component Gaussian components per state.

4.3 Recognition Results

There are two test sets. The first is "HTest04" collected by HKUST and released in 2005, which comprises of 4 hours of data with 24 phone calls. The second is "GDTest", comprised of half hour of self-collected data with 354 conversations by phone.

The character error rate (CER) is used to estimate the recognition performance, which is obtained by a one-pass decoding [10]. The recognition results are shown in Tab. 1. The first row is the result of baseline using a canonical lexicon. The second rows show the results of MPEPM without AM co-training, while the last row is the result of MPEPM with AM co-training. From these results, MPEPM shows its effectiveness.

5. Conclusions and Discussion

In this work, we presented a discriminative pronunciation modeling method based on the MPE training. We rewrote the auxiliary function of the MPE training at word level, and incorporated PMs into it. By doing this, we explored a way to discriminatively co-train the acoustic models and the pronunciation models in an iterative fashion. We demonstrated that the required statistics could be obtained in the standard MPE training. Thus, this method is easy and efficient to implement. Finally, experimental results on Mandarin conversational speech recognition task demonstrated the effectiveness of this method.

Since the current state-of-the-art systems make use of Deep Neural Networks (DNNs), we would like to discuss the possibilities of this MPEPM to be used in DNNs based framework. Currently, there are two main approaches to incorporate DNNs in acoustic modeling: the TANDEM system and the hybrid system. For TANDEM system, DNNs act as feature extractors to derive bottleneck features, which can be used to train traditional HMM/GMM. Thus, this MPEPM implementation keeps constant. For the hybrid system, DNNs estimate posterior probabilities of states of HMMs. This MPEPM can be efficiently implemented within the sequence-discriminative training of DNNs, as they are all based on reference and hypothesis lattices, especially for the one using the state-level minimum Bayes risk (sMBR) criterion, which is derived from the MPE criterion.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Nos. 11161140319, 91120001, 61271426), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA06030100, XDA06030500), the National 863 Program (No. 2012AA012503) and the CAS Priority Deployment Project (No. KGZD-EW-103-2).

References

[1] O. Vinyals, L. Deng, D. Yu, and A. Acero, "Discriminative pronunciation learning using phonetic decoder and minimum classication-error criterion," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009, Apr., pp. 4445-4448.

[2] I. Badr, I. McGraw, and J. Glass, "Pronunciation learning from continuous speech, Proc. Interspeech, Florence, 2011.

[3] I. McGraw, I. Badr, and J. Glass, "Learning lexicons from speech using a pronunciation mixture model," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 2, pp. 357-366, Feb. 2013.

[4] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion, Speech Communication, vol. 50, no. 5, pp. 434-451, 2008.

[5] X. Lei, W. Wang, and A. Stolcke, "Data-driven lexicon expansion for mandarin broadcast news and conversation speech recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009, Apr., pp. 4329-4332.

[6] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Ph. D. thesis, Cambridge University, 2003.

[7] H. Schramm and X. Aubert, "Efficient integration of multiple pronunciations in a large vocabulary decoder," in 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000. ICASSP 2000. Proceedings, vol. 3, pp. 1659-1662 vol. 3.

[8] N. Kumar, "Investigation of silicon auditory models and generalization of linear discriminant analysis for improved speech recognition," Ph.D. thesis, Johns Hopkins University, Baltimore, 1997.

[9] C. Liu and Y. Yan, "Robust state clustering using phonetic decision trees," Speech Communication, vol. 42, no. 3-4, pp. 391-408, 2004.

[10] J. Shao, T. Li, Q. Zhang, Q. Zhao and Y. Yan, "A One-Pass Real-Time Decoder Using Memory-Efficient State Network," IEICE TRANSACTIONS on Information and Systems, Vol. E91-D, No. 3, pp. 529-537, Mar. 2008.