DISCRIMINATIVE INCORPORATION OF EXPLICITLY TRAINED TONE MODELS INTO LATTICE BASED RESCORING FOR MANDARIN SPEECH RECOGNITION

Hao Huang and Jie Zhu

Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, 200240, P.R. China.
{haohuang, zhujie}@sjtu.edu.cn

ABSTRACT

Explicit tone modeling has been widely discussed in recent Mandarin speech recognition research. In this paper, a discriminative method of incorporating explicitly trained tone models into lattice based rescoring is proposed. The method is to use discriminative trained model weights to scale the acoustic model and tone model distributions. The weights are trained by the Minimum Phone Error using the Extended Baum Welch algorithm. To take into account different phonetic contexts, various model weighting schemes are evaluated. A smoothing technique is introduced to make model weight training more robust to over fitting. The proposed method is evaluated on tonal syllable output speech recognition tasks on a Mandarin LVCSR database. Results show the proposed method has achieved significant error reduction than traditional global weight approach. Comparison with the traditional embedded tone modeling is also made, which shows the importance of the proposed method when explicit tone modeling approach is applied.

Index Terms— explicit tone model incorporation, minimum phone error, discriminative training, Mandarin speech recognition.

1. INTRODUCTION

Tone plays an important role in reducing ambiguity in Mandarin speech recognition. Utilization of tone information has proved to be successful in improving accuracy in large vocabulary speech recognition (LVCSR). Most state-of-the-art Mandarin speech recognition systems adopt embedded tone modeling approach, where the F_0-related tonal features are appended into one single stream with the traditional spectral feature (Mel Frequency Cepstral Coefficient, MFCC/ Perceptual Linear Prediction, PLP) for training and recognition, as has been done in [1]. Explicit tone modeling is to build tone classifier separately and add tone scores from the classifier in lattice rescoring. Explicit tone modeling can exploit supra-segmental nature of the tones or use better tone classifier, which can lead to better recognition performance than the embedded approach, as shown in [2]. Some work adopted a hybrid framework, i.e., generate lattices by using the embedded approach and then use improved tone models to obtain further improvement [3].

Therefore, the two major tasks for explicit tone modeling are to build a tone classifier for better discriminating the tones and integrate tone scores from the classifier into LVCSR to improve the performance of continuous speech recognition. For the first task, many explicit tone modeling techniques have been studied in last two decades. In recent works, several tone models have been reported, including the overlapped ditone Gaussian mixture model [4], the decision tree based tone model with polynomial regression coefficient features [5], and support vector machines [6]. All have shown significant tone recognition improvement by taking advantages of supra-segmental nature of the tones.

After the explicitly trained tone models are obtained, how to find an optimal integration of the tone models into second pass rescoring is another important issue for tone problem solving in Mandarin speech recognition. Traditional approach is to use global acoustic score weight and tone score weight (commonly obtained by heuristics or a grid search process) to scale the probabilities, which might not lead to the best result. This is because the acoustic model and tone models are independently trained and need a better interpolation. On the other hand, the global weight could not take into account the local phonetic/semantic scenario, such as what exact a phone or word has been uttered. In [7], the word level prosody model was proposed which is capable of grasping tone articulation variation within a word. In this paper, we propose discriminative model weight training (DMWT) for obtaining better model dependent weights to scale the acoustic scores and tone scores. The weights are trained using the Extend Baum-Welch (EBW) algorithm by the recently popular Minimum Phone Error (MPE) criterion [8, 9]. We evaluated several model weighting schemes: tonal syllable dependent, final model dependent, and model combination dependent. In [7], the less frequent words, tonal syllable-dependent tone model or plain tone models are used as back off. We propose a similar technique: smoothing between the weights derived from different weighting schemes is adopted to eliminate overtraining. The tonal syllable output speech recognition tasks are performed to evaluate DMWT. Experiments show a 3.3% error reduction (9.5% relative) has been achieved by DMWT

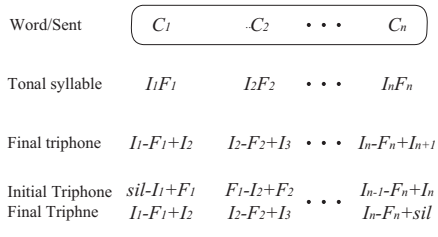

Fig. 1. scheme of DMWT model combination

than that with global model weight integration scheme. It is also shown DMWT helps gain further improvement to the embedded tone modeling approach.

The remainder of this paper is organized as follows: In Section 2, tone model incorporation framework in lattice rescoring is described. The weight combination scheme is described. Section 3 gives an overview of the MPE objection function and the weight updating equation using EBW is derived. Section 4 presents the experimental results. Finally in section 5 the conclusions drawn from the work are given.

2. TONE MODEL INCORPORATION FRAMEWORK FOR LATTICE RESCORING

2.1. Tone model incorporation

The total score for an arc in the lattice is computed based on scores from the parallel models:

$$\begin{aligned}&\psi(q) = \textstyle\sum_i^I \eta_i \psi_i(q) \\ &= \eta_A \alpha \psi_{AM}(q) + \eta_T \beta \psi_{TM}(q) + \psi_{LM}(q) + \psi_{WP}\end{aligned} \tag{1}$$

where $\psi(q)$ is the score from the ith parallel model. $\psi_{AM}(q)$ is the acoustic model (AM) score and $\psi_{TM}(q)$ is the tone model (TM) score. ψ_{LM} is the language model score and ψ_{WP} is the word penalty. α and β are respectively the global factor for the acoustic and tone scores and are selected empirically. η_A and η_T are respectively the acoustic and tone probability weight which are to be trained. We denote $\eta = (\eta_A, \eta_T)$ as a model weight pair, where $\eta_A > 0, \eta_T > 0$ and $\eta_A + \eta_T = 1$ are assumed.

2.2. Acoustic modeling for Mandarin speech recognition and weight combination scheme

Fig.1 demonstrates the triphone based modeling structure of a multi-character Chinese word/sentence (silence is assumed at both the beginning and the end). As shown, each character in the word can be pronounced as a tonal syllable. And each tonal syllable can be divided into two parts: initial (I) and final (F). Each part can be modeled by a triphone (initial triphone or final triphone), according to its context. Hence, we evaluated the following weighting schemes:

- Tone syllable dependent (*DMWT_TSD*): Associate a unique model weight pair with each tonal syllable. DMWT_TSD considers the initial and final type of the syllable;
- Final model dependent (*DMWT_FMD*): Associate a unique model weight pair with each final triphone. This scheme takes into account the initial type of the following syllable;
- Model combination dependent (*DMWT_MCD*): Associate a unique weight pair with each different initial-final triphone combination. The scheme is capable of taking into account the final type of the preceding syllable.

3. DISCRIMINATIVE MODEL WEIGHT TRAINING

3.1. MPE objective function

The model weights are trained according to the MPE objective function. Given a training set of acoustic observation sequences $\mathcal{O} = \{\mathcal{O}_1, ..., \mathcal{O}_r, ..., \mathcal{O}_R\}$, the MPE criterion is to minimize the average phone error of the observation sequences [8, 9]:

$$\mathcal{F}_{\text{MPE}} = \sum_r^R \frac{\sum_{s \in S} P(\mathcal{O}_r|s)^\kappa P(s)^\kappa Acc(s, s_r)}{\sum_{s' \in S} P(\mathcal{O}_r|s')^\kappa P(s')^\kappa} \tag{2}$$

where $P(\mathcal{O}_r|s)$ is the acoustic score for sentence s and $P(s)$ is the language model. κ is a scaling factor for reducing dynamic range for acoustic scores. $Acc(s, s_r)$ is the raw phone accuracy for hypothesis s and can be calculated in terms of the sum of the accuracy of each arc contained in s. More details of MPE training can be found in [8, 9].

3.2. Extended Baum Welch model weight optimization

When the model weights are to be trained, the MPE objective maximization is accomplished with the EBW algorithm [10] when satisfying the positive and sum-to-one conditions:

$$\eta'_{m,i} = \frac{\eta_{m,i} \left(\partial \mathcal{F}_{\text{MPE}} / \partial \eta_{m,i}|_\eta + C\right)}{\sum_i \eta_{m,i} \left(\partial \mathcal{F}_{\text{MPE}} / \partial \eta_{m,i}|_\eta + C\right)}, \tag{3}$$

where $\eta_{m,i}$ and $\eta'_{m,i}$ are respectively current and newly estimated weights for the ith model score in pair m. C is a constant used to ensure positive probability weight. The differential of $\mathcal{F}_{\text{MPE}}$ w.r.t certain model weight $\eta_{m,i}$ needs to be computed. According to the chain rule:

$$\frac{\partial \mathcal{F}_{\text{MPE}}}{\partial \eta_{i,m}} = \frac{\partial \mathcal{F}_{\text{MPE}}}{\partial \psi(q)} \frac{\partial \psi(q)}{\partial \eta_{i,m}}. \tag{4}$$

The first item is can be computed by [9]:

$$\partial \mathcal{F}_{\text{MPE}} / \partial \psi(q) = \kappa \gamma_q^{\text{MPE}}, \tag{5}$$

where $\gamma_q^{\text{MPE}} = \gamma_q \left(c(q) - c_{avg}\right)$. γ_q is the posterior probability of passing arc q. $c(q)$ is the average phone accuracy for all the sentence hypothesis that contains arc q and c_{avg} is the average accuracy of all the hypothesis in the lattice. More details about computation of these statistics can be found in [8, 9]. The second item in Eqn. (4) is computed by:

$$\partial \psi(q) / \partial \eta_i = \psi_i(q), \tag{6}$$

AM	TM	Weight	TSER (%)	ERR (%)
MSR	N	–	48.7	–
MSR	Y	global	41.3	15.1
MPE	N	–	40.9	16.0
MPE	Y	global	34.8	28.5
MPE	Y	TSD	34.1	30.0
MPE	Y	FMD	32.9	32.4
MPE	Y	MCD	32.5	33.2
MPE	Y	smoothing	**31.5**	35.3

Table 1. Recognition results for tonal syllable output tasks

Iter.	E=150	E=200	E=250	E=300
0	34.8	34.8	34.8	34.8
1	34.2	34.3	34.5	34.5
2	34.2	34.1	34.4	34.4
3	34.3	34.1	**34.1**	34.3
4	34.4	34.2	34.1	34.3

(a) DMWT TSD

Iter.	E=150	E=200	E=250	E=300
0	34.8	34.8	34.8	34.8
1	33.4	33.6	33.6	33.7
2	33.1	33.1	33.1	33.1
3	33.1	32.9	**32.9**	32.9
4	33.1	33.1	33.0	32.9

(b) DMWT FMD

Iter.	E=150	E=200	E=250	E=300
0	34.8	34.8	34.8	34.8
1	32.7	33.0	33.2	34.4
2	32.8	32.7	32.5	32.6
3	33.1	32.7	32.5	32.5
4	33.3	33.0	**32.5**	32.6

(c) DMWT MCD

Table 2. TSER with different weighting scheme

which is the score from the ith parallel model. Then the iterative updating function for DMWT can be written as:

$$\eta'_{m,i} = \frac{\kappa\gamma_q^{\mathrm{MPE}}\eta_{m,i}\psi_i(q)|_\eta + C\eta_{m,i}}{\sum_i \left(\kappa\gamma_q^{\mathrm{MPE}}\eta_{m,i}\psi_i(q)|_\eta + C\eta_{m,i}\right)}. \quad (7)$$

4. EXPERIMENTS AND RESULTS

4.1. Database and configurations

The experiments are performed on a Mandarin LVCSR database. The corpus from Microsoft Research Asia [11] is used for training, which contains read speech of 31.5 hours from 100 male students, for a total 19 688 utterances and 454 291 tonal syllables. In the testing phase, the test uses

Iter.	E=150	E=200	E=250	E=300
0	0.433	0.433	0.433	0.433
1	0.427	0.428	0.428	0.429
2	0.424	0.424	0.426	0.427
3	0.421	0.423	0.424	0.425
4	0.420	0.422	0.423	0.423

(a) DMWT TSD

Iter.	E=150	E=200	E=250	E=300
0	0.433	0.433	0.433	0.433
1	0.413	0.416	0.418	0.420
2	0.400	0.405	0.408	0.411
3	0.392	0.397	0.401	0.404
4	0.387	0.392	0.396	0.399

(b) DMWT FMD

Iter.	E=150	E=200	E=250	E=300
0	0.433	0.433	0.433	0.433
1	0.375	0.386	0.393	0.399
2	0.343	0.356	0.366	0.376
3	0.324	0.336	0.346	0.355
4	0.312	0.322	0.331	0.341

(c) DMWT MCD

Table 3. Expected error rates for DMWT iterations

additional 0.74 hour 500 utterances (9 570 syllables) from another 25 male speakers. Speech waveforms are sampled at 16bit and 16 kHz. Each frame of the acoustic front-end is represented by a 39 dimensional vector, consisting of 12 MFCCs and normalized log energy and their delta and acceleration. Our tone model is based on the hidden conditional random fields [12], which has shown a slight tone error rate (TER) reduction than EBW trained HMMs, when using the same structure and observations (including normalized F_0 and ΔF_0, the normalized log energy and its first and second derivative). TER on the test data is 28.7%. Conditional probabilities are used as tone scores in Eqn. (1).

4.2. Experimental results

The proposed DMWT is evaluated on tonal syllable output speech recognition task provided by the MSR toolbox in [11]. Because we focus on the acoustics, no language model is used. As mentioned in [5], measuring the tonal syllable recognition performance is a good evaluation of the acoustic model resolution of a recognizer, because it is done purely at the phonetic level by removing the language model from the LVCSR decoding process. Recognition is carried out in two passes. The first pass is a normal time-synchronous beam search with the acoustic model and the output is a tonal syllable lattice. The second pass is to rescore within the lattice including the acoustic and tone models to find the most likely tonal syllable sequences. We first show direct integration of

Model	ML	MPE
TSER	41.8	35.5

Table 4. Recognition results for embedded approach

tone model without weight training. Global setting in Eqn.(1) is $\alpha = 1$, $\beta = 4.5$ and $\psi_{WP} = 35$. It can be seen in upper part of Table 1 when the TMs are incorporated into the MSR baseline (trained with ML) and MPE trained acoustic model, TSER is reduced from 48.3% to 41.3% and from 40.9% to 34.8%.

Then we experiment with DMWT. Smoothing constant C in Eqn. (7) can be set to $C = E\sum_i \left|\kappa\gamma_q^{\mathrm{MPE}}\eta_{m,i}\psi_i(q)|_\eta\right|$ where E is a constant to control the training speed and convergence. Results and DMWT iterations with different constant E are evaluated in Table 2. As shown, for the scheme of DMWT_TSD, DMWT_FMD, DMWT_MCD, TSER is considerably reduced from 34.8% achieved by global weight to 34.1%, 32.9% and 32.5%. In the training phase, the number of trained weights for the three weighting schemes is 1 232, 40 243 and 1 806 565. Those unseen weights in the test set will be given the default global value. The DMWT_MCD is consistently better than DMWT_FMD with an absolute gain of 0.4%. From the results it can be seen performance is improved when the number of weight pairs increases.

Table 3 demonstrates the expected error rate by evaluating MPE objective function of DMWT with different smoothing constants. We can see the expected error reduction of DMWT_MCD is much larger than that of DMWT_FMD. However, improvement of DMWT_MCD to DMWT_FMD is marginal. It is shown in Table 2 recognition accuracy peaks at certain iteration and degrades afterwards. This is mainly because DMWT_MCD is liable to overtraining for there is far less training data per model weight pair than in TSD and FMD (The average number of training samples per model weight pair is 21 436, 656 and 15 for TSD, FMD and MCD). To make DMWT_FMC more robust, the model weights are smoothed using interpolation between FMD and MCD derived weights, i.e. $\boldsymbol{\eta}_{smooth} = \rho\boldsymbol{\eta}_{\mathrm{FMD}} + (1-\rho)\boldsymbol{\eta}_{\mathrm{MCD}}$ and $\rho = 0.35$ has shown to obtain the best result. This is denoted as smoothing shown in Table 1, which leads to about 1.0% further improvement.

4.3. Comparisons with the embedded approach

To compare the results of two tone modeling approach, we have also evaluated performance of the embedded tone modeling: the ML and MPE trained acoustic model using standard 39-dim MFCC plus normalized F_0 and ΔF_0 tonal features with interpolated F_0 in unvoiced part of a syllable. The TSERs are shown in Table 4. By comparing the results with those in Table 1 and Table 4, we can see the explicit tone modeling using global weight only outperform the embedded approach by a slight margin. We think it might because of the inferior TM we have exploited. It is shown when DMWT is applied, explicit approach is significantly better than the embedded tone modeling. From above we believe the proposed DMWT is essential to obtain an optimal result when utilizing explicit tone modeling in lattice rescoring.

5. CONCLUSION

We have shown the use of discriminative model weight training (DMWT) when incorporating the explicitly trained tone models into lattice based rescoring. A smoothing technique for reducing overtraining is presented and introduced further improvement. Results on tonal syllable output speech recognition tasks have shown DMWT derived model weights significantly outperformed the global model weight scheme. The proposed DMWT also provides a promising framework for optimal fusion of heterogeneous features or models in lattice rescoring. We believe it is also applicable to other explicit tone modeling techniques as we have mentioned. Future work includes resorting to better tone classifier and performing character output speech recognition tasks to evaluate the effectiveness of DMWT.

6. REFERENCES

[1] C. H. Huang, Side F, "Pitch tracking and tone features for mandarin speech recognition," in *Proc. of ICASSP*, 2000. 1523-1526.

[2] X. Lei, M. H. Siu, M. Hwang, M. Ostendorf, et al, "Improved Tone Modeling for Mandarin Broadcast News Speech Recognition," in *Proc. of Interspeech*. 2006. 1277-1280.

[3] H. L. Wang, Y. Qian , F. K. Soong, J. L. Zhou, et al, "Improved Mandarin Speech Recognition by Lattice Rescoring with Enhanced Tone models," in *Proc. of ISCSLP*, pp. 445-443, 2006.

[4] Y. Qian, T. LEE, Y. J. Li, "Overlapped ditone modeling for tone recognition in continuous Cantonese speech," in *Proc. of Eurospeech*, 2003. 1845-1848.

[5] P. F. Wong, M. H. Siu, "Decision tree based tone modeling for Chinese speech recognition," in *Proc. of ICASSP*. 2004. 905-908.

[6] G. Peng, W. S. Wang, "Tone recognition of continuous Cantonese speech based on support vector machines," *Speech Communication*, 2005, 45: 49-62.

[7] X. Lei, M. Ostendorf, "Word level tone modeling for Mandarin Speech Recognition," in *Proc. of ICASSP*, 2007. IV-665-668.

[8] D. Povey, P. C. Woodland, "Minimum Phone Error and I-smoothing for Improved Discriminative Training," in *Proc. of ICASSP,* 2002. 105-108.

[9] Povey D. *Discriminative Training for Large Vocabulary Speech Recognition*, Ph.D. thesis, Cambridge University, 2004.

[10] P. S. Gopalakrishnan, D. Kanevsky, A. Nadas, et al. "A generalization of the Baum algorithm to rational objective functions," in *Proc. of ICASSP*, 1989. 631-634.

[11] E. Chang, Y. Shi, J. L. Zhou, et al, "Speech lab in a box: a Mandarin speech toolbox to jumpstart speech related research," in *Proc. of Eurospeech*, 2001. 2779-2782.

[12] A. Gunawardana, M. Hahajan, A. Acero A, J. C. Platt, "Hidden conditional random fields for phone classification," in *Proc. of Eurospeech*, 2005. 1117-1120.