Discriminative tone model training and optimal integration for Mandarin speech recognition

Huang Hao Zhu Jie

(Department of Electronic Engineering, Shanghai Jiaotong University, Shanghai 200240, China)

Abstract: Two discriminative methods for solving tone problems in Mandarin speech recognition are presented. First, discriminative training on the HMM (hidden Markov model) based tone models is proposed. Then an integration technique of tone models into a large vocabulary continuous speech recognition system is presented. Discriminative model weight training based on minimum phone error criteria is adopted aiming at optimal integration of the tone models. The extended Baum Welch algorithm is applied to find the model-dependent weights to scale the acoustic scores and tone scores. Experimental results show that tone recognition rates and continuous speech recognition accuracy can be improved by the discriminatively trained tone model. Performance of a large vocabulary continuous Mandarin speech recognition system can be further enhanced by the discriminatively trained weight combinations due to a better interpolation of the given models.

Key words: discriminative training; minimum phone error; tone modeling; Mandarin speech recognition

Tone recognition is an important task for Mandarin speech recognition due to the tonal nature of the language. There has been much work done discussing tone modeling to improve tone recognition accuracy. The most popularly applied is the frame and the HMM based approach[1]. In addition, other approaches such as the stochastic polynomial tone model (SPTM) proposed in Ref. [2] and the decision tree based tone model[3] have also been proposed.

In state-of-the-art speech recognition systems, discriminative training is commonly employed to obtain the best recognition accuracy. Among the discriminative criteria are the maximum mutual information[4], the minimum classification error (MCE)[5] and the minimum phone error (MPE)[6–7]. The recently proposed MPE is currently popular for speech recognition and can significantly reduce word error rate. The MPE objective function is an approximation of phone accuracy directly related to the recognition result. The extended Baum Welch (EB) update equations used in MPE training have the advantage of being simple to implement as they do not require statistics from more than one iteration of training[7].

We focus on the HMM-based approach and investigate discriminative training of the HMM based tone model which is referred to as the minimum tonal error (MTE). After obtaining the discriminatively trained tone models, we propose a method to discriminatively integrate tone information into the existing systems. In previous work on tone modeling incorporation such as in Refs. [2 – 3], a global acoustic and tone model weight is commonly applied which may not obtain an optimal result. Liu et al. [8] proposed an MCE based stream weight optimization for audio-visual LVSCR. Inspired by this work, we propose discriminative training on model weight using the EB algorithm under the MPE criterion. Results evaluated based on tone classification and large vocabulary continuous speech recognition tasks indicate the effectiveness of both proposed methods.

1 Minimum Tonal Error Training

Let $\boldsymbol{\Gamma} = \{\Gamma_i\}_{i=1}^{5}$ be the 5 tone classes and $\boldsymbol{\Gamma} = \{\Gamma_{i,j}\}_{j=1}^{J_i}$ be the J_i sub models of class Γ_i. Given R tonal syllables, let $O_r\{r = 1, 2, \ldots, R\}$ be the observation frames for syllable r. The MTE objective function is an approximation of tone recognition accuracy and can be expressed as a sum of tone accuracy weighted by model posterior probability:

$$F_{\mathrm{MTE}}(\boldsymbol{\Gamma}) = \sum_{r}^{R} \sum_{i=1, j=1}^{5, J_i} P(\Gamma_{i,j} \mid O_r) \mathrm{Acc}(\Gamma_{i,j}, \Gamma_r) \quad (1)$$

where $P(\Gamma_{i,j} \mid O_r)$ is the posterior probability from tone model $\Gamma_{i,j}$; Acc$(\cdot)$ is the accuracy measure and we have $A(\Gamma_{i,j}, \Gamma_r) = 1$ if $\Gamma_{i,j}$ is the correct tone Γ_r from transcription and 0 otherwise (There might be a different model belonging to the tone class). The objective of Eq. (1) can be written as

Received 2006-11-06.

Biographies: Huang Hao (1976—), male, graduate; Zhu Jie (corresponding author), male, doctor, professor, zhujie@ sjtu. edu. cn.

$$F_{\text{MTE}}(\boldsymbol{\Gamma}) = \sum_{r}^{R} \sum_{1}^{5} \frac{P(O \mid \Gamma_{i,j})^{\kappa^{\text{MTE}}} \text{Acc}(\Gamma_{i,j}, \Gamma_r)}{\sum_{i}^{5} \sum_{j}^{J_i} P(O \mid \Gamma_{i,j})^{\kappa^{\text{MTE}}}} \tag{2}$$

where κ^{MTE} is a scaling factor for reducing the dynamic range to prevent domination by one model. To maximize the objective function, an optimization method using the extend Baum Welch algorithm is used. Tone model parameters can be re-estimated iteratively based upon parameters of previous iterations:

$$\mu'_{ijkm} = \frac{\{\theta_{ijkm}^{\text{num}}(O) - \theta_{ijkm}^{\text{den}}(O)\} + D_{ijkm}\mu_{ijkm}}{\{\gamma_{ijkm}^{\text{num}} - \gamma_{ijkm}^{\text{den}}\} + D_{ijkm}} \tag{3}$$

$$\sigma^{2\prime}_{ijkm} = \frac{\{\theta_{ijkm}^{\text{num}}(O^2) - \theta_{ijkm}^{\text{den}}(O^2)\} + D_{ijkm}(\sigma_{ijkm}^2 + \mu_{ijkm}^2)}{\{\gamma_{ijkm}^{\text{num}} - \gamma_{ijkm}^{\text{den}}\} + D_{ijkm}} \tag{4}$$

where μ_{ijkm}, σ^2_{ijkm}, σ'_{ijkm} and, $\sigma^{2\prime}_{ijkm}$ are current and newly estimated means and variances of Gaussian mixture m in state k of model $\Gamma_{i,j}$; and D_{ijkm} is the positive smoothing constant. Several statistics in Eqs. (3) and (4) need to be calculated before updating the model parameters:

$$\gamma_{ijkm}^{\text{num}} = \sum_{r}^{R} \sum_{t=1}^{T_r} \gamma_{ijkm}(t) \max(0, \gamma_{i,j}^{\text{MTE}}) \tag{5}$$

where $\gamma_{ijkm}(t)$ is the within model Gaussian posterior occupancy of the mixture component m in state k for tone model $\Gamma_{i,j}$ at time t and can be calculated through a forward-backward pass; $\gamma_{i,j}^{\text{MTE}} = \gamma_{i,j}(\text{Acc}(\Gamma_{i,j}) - c_{\text{avg}}^{\text{MTE}})$, $\gamma_{i,j}$ is the tone model posterior probability from $\Gamma_{i,j}$; $c_{\text{avg}}^{\text{MTE}}$ is the average tone accuracy for all models, i. e.,

$$c_{\text{avg}}^{\text{MTE}} = \sum_{i=1,j=1}^{5,J_i} P^{\kappa^{\text{MTE}}}(\Gamma_{i,j} \mid O_r) \text{Acc}(\Gamma_{i,j}, \Gamma_r) \tag{6}$$

For numerators, the sum-of-data $\theta_{ijkm}^{\text{num}}(O)$ and sum-of-squared-data $\theta_{ijkm}^{\text{num}}(O^2)$ in Eq. (3) and (4) are listed as follows:

$$\theta_{ijkm}^{\text{num}}(O) = \sum_{r}^{R} \sum_{t=1}^{T_r} \gamma_{ijkm}(t) \max(0, \gamma_{i,j}^{\text{MTE}}) O(t) \tag{7}$$

$$\theta_{ijkm}^{\text{num}}(O^2) = \sum_{r}^{R} \sum_{t=1}^{T_r} \gamma_{ijkm}(t) \max(0, \gamma_{i,j}^{\text{MTE}}) O^2(t) \tag{8}$$

The $\gamma_{ijkm}^{\text{den}}$ in Eqs. (3) and (4), the sum-of-data and sum-of-squared-data for denominators $\theta_{ijkm}^{\text{den}}(O)$ and $\theta_{ijkm}^{\text{den}}(O^2)$ can be obtained by replacing $\gamma_{i,j}^{\text{MTE}}$ with $-\gamma_{i,j}^{\text{MTE}}$ in Eqs. (5), (7) and (8). Transition probability and mixture weight have similar forms as proposed in the MPE/MWE training and will not be discussed redundantly. More details of these statistics can be found in Refs. [6 – 7]. From the above we can see that the underlying characteristics of MTE are similar to the MPE based approach. Those hypotheses from the right tone will have higher accuracy than average and provide positive contributions, weighted in proportion to the tone model posterior likelihoods and the differences between tone accuracy and average accuracy[9].

2 MPE Based Model Probability Weight Training

2.1 Tone modeling integration framework for LVCSR

The tone modeling integration framework can be expressed as

$$P(O \mid q) = \prod_{i=1}^{I} P(O_i \mid \xi_i)^{\eta_i} \tag{9}$$

where $P(O_i \mid \xi_i)$ is the i-th model probability from model ξ_i, and η_i is the model probability weight for ξ_i. Eq. (9) can be rewritten as

$$P(O \mid q) = P(O_{\text{A}} \mid \lambda_q)^{\eta_1} P_{\Gamma}(O_{\text{T}} \mid \Gamma)^{\eta_2} p_{\text{L}}(l \mid L)^{\eta_3} \tag{10}$$

where O_{A} are acoustic feature observations for arc q and $P(O_{\text{A}} \mid \lambda_q)$ is the probability from acoustic model λ_q; O_{T} is tone related feature for the segment; $P(O_{\text{T}} \mid \Gamma)$ is the tone model probability. $P_{\text{L}}(\cdot)$ is the GMM based duration model in Ref. [2]. The likelihood of the length l generated by the tone pattern $\Gamma_{i,j}$ can be expressed as

$$p_{\text{L}}(l \mid \Gamma_{i,j}) = \sum_{m=1}^{M} w_{ijm}^{L} N(l - \mu_{ijm}^{L}, \sigma^{2L}_{ijm}) \tag{11}$$

where w_{ijm}^{L}, μ_{ijm}^{L} and σ_{ijm}^{L} are the weight coefficients, means and variances of mixture component m for model j of tone i. In Eq. (9), η_1, η_2 and η_3 are the model weights which are to be trained as discussed in later part of this section.

Since the acoustic models and tone models are separately trained, the weighted distributions are essential in obtaining an optimal result. And another reason for such integration lies in the diversity of the models, i. e., there are different tonal modeling techniques other than frame-based HMMs. Therefore, the generalized framework of tone model integration in Eq. (9) is reasonable when these heterogeneous models are integrated into continuous speech recognition.

2.2 MPE objective function

The model weights are trained according to the MPE objective function. Given a training set of observation sequences $O = \{O_1, \ldots, O_u, \ldots, O_U\}$, the MPE criterion for acoustic modeling is to minimize the average phone error of the observation sequences using the objective[6-7]:

$$F_{\text{MPE}}(\boldsymbol{\lambda}, \boldsymbol{\Gamma}, \boldsymbol{\eta}) = \sum_{u}^{U} \frac{\sum_{s \in S} P(O_u \mid s)^{\kappa^{\text{MPE}}} P(s)^{\kappa^{\text{MPE}}} \text{Acc}(s)}{\sum_{s' \in S} P(O_u \mid s)^{\kappa^{\text{MPE}}} P(s)^{\kappa^{\text{MPE}}}} \tag{12}$$

where $\boldsymbol{\lambda}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\eta}$ are the acoustic model, the tone model and the model dependent weight matrix, respectively; $P(O_u \mid s)$ is the acoustic score (including the tone score) for sentence s and $P(s)$ is the language model; κ^{MPE} is a scaling factor for reducing the dynamic range for acoustic scores; $\text{Acc}(s)$ 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:

$$\text{Acc}(s) = \sum_{q \in S} \text{Acc}(q) \tag{13}$$

where Acc (q) is phone arc accuracy and defined as[6-7]

$$\text{Acc}(q) = \max_{z} \begin{cases} -1 + 2e(q, z) & \text{if } z \text{ and } q \text{ are the same phone} \\ -1 + e(q, z) & \text{if } z \text{ and } q \text{ are different phones} \end{cases}$$

where $e(q, z)$ is the ratio of overlapped lengths to the transcription length of arc z. More details of the MPE calculation can be found in Refs. [6 − 7].

2. 3 Extended Baum Welch model weight optimization

Denote the model dependent weights as $\eta = [\eta_{b,i}]_{B \times I}$, where $\eta_{b,i}$ is the i-th model weight for model combination b. B is the number of total possible model combinations and I is the number of models for each combination. In order to satisfy positive and sum-to-one conditions, the probability weight can be optimized with the EB algorithm[10]:

$$\eta'_{b,i} = \frac{\eta_{b,i}\left(\left.\dfrac{\partial F_{\text{MPE}}(\boldsymbol{\lambda}, \boldsymbol{\Gamma}, \boldsymbol{\eta})}{\partial \eta_{b,i}}\right|_{\eta} + C \right)}{\sum_{i} \eta_{b,i}\left(\left.\dfrac{\partial F_{\text{MPE}}(\boldsymbol{\lambda}, \boldsymbol{\Gamma}, \boldsymbol{\eta})}{\partial \eta_{b,i}}\right|_{\eta} + C \right)} \tag{14}$$

where $\eta_{b,i}$ and $\eta'_{b,i}$ are the current and the newly estimated model weights, respectively; C is a constant used to ensure positive probability weight. According to the chain rule, the differential of F_{MPE} w. r. t a certain model weight in (14) can be calculated as

$$\frac{\partial F_{\text{MPE}}(\boldsymbol{\lambda}, \boldsymbol{\Gamma}, \boldsymbol{\eta})}{\partial \eta_i} = \frac{\partial F_{\text{MPE}}(\boldsymbol{\lambda}, \boldsymbol{\Gamma}, \boldsymbol{\eta})}{\partial \log(O \mid q)} \frac{\partial \log(O \mid q)}{\partial \eta_i} \tag{15}$$

The first item can be computed as

$$\frac{\partial F_{\text{MPE}}(\lambda, \Gamma, \eta)}{\partial \log(O \mid q)} = \kappa^{\text{MPE}} \gamma_q^{\text{MPE}} \tag{16}$$

where $\gamma_q^{\text{MPE}} = \gamma_q(c(q) - c_{\text{avg}}^{\text{MPE}})$, γ_q is the posterior probability of passing arc q, $c(q)$ is the average phone accuracy for all of the sentence hypothesis that contains arc q, and $c_{\text{avg}}^{\text{MPE}}$ is the average accuracy of all the hypothesis in the lattice. The second item in Eq. (15) is computed by $\partial \log(O \mid q) / \partial \eta_i = \log(O_i \mid \xi_i)$. By substitution and rearrangement, the iterative updating function for weight training can be written as

$$\eta'_{b,i} = \frac{\kappa^{\text{MPE}} \gamma_q^{\text{MPE}} \eta_{b,i} \log(O_i \mid \xi_i) \mid_{\eta} + C\eta_{b,i}}{\sum_{i} (\kappa^{\text{MPE}} \gamma_q^{\text{MPE}} \eta_{b,i} \log(O_i \mid \xi_i) \mid_{\eta} + C\eta_{b,i})} \tag{17}$$

3 Experiments

3. 1 Front-end configurations

The acoustic feature of each frame is represented by a 39 dimensional feature vector, consisting of 12 MFCCs and normalized log energy and their delta and acceleration. The tonal speech feature includes the normalized log energy and its first and second derivatives, the F0 and its first derivative. The pitch detection algorithm is sub-harmonic summation (SHS) based with a dynamic programming technique to remove pitch errors[1].

3. 2 Database

The experiments are performed on a large vocabulary continuous Mandarin speech recognition database. The corpus from Microsoft Research Asia[11] is used for training. The database contains read speech of about 31. 5 h from 100 male students, for a total of 19 688 utterances and 454 294 tonal syllables. In the testing phase, the MSR[11] test uses an additional 0. 74 h 500 utterances (9 570 syllables in total) from another 25 male speakers. Speech waveforms are sampled at 16 bit and 16 kHz.

The baseline of continuous speech recognition system uses context dependent triphone units for modeling the Mandarin tonal syllables. After state-tying there are 2 392 tied states. Each state is composed of 8 Gaussian mixtures. For tone modeling, each HMM has 3 emitting states with 16 Gaussians per state.

3. 3 Results

The first set of experiments is conducted to classify the tonal syllables in the test utterances to evaluate the effectiveness of MTE training. We performed the tone classification task with models trained by different methods. Tab. 2 demonstrates the recognition results. The ML _ CI, ML _ CD are respectively the 5-context independent (CI) HMM tone model and the 23-context dependent (CD) tone model described in Ref. [12] trained with maximum likelihood (ML). We have done MTE training from different initial models of ML _ CI to MTE _ CI and from ML _ CD to MTE _ CD, respectively. The scale factor in Eqs. (1)

and (2) is empirically selected with 18, and constant D in Eqs. (3) and (4) is set to $D = E\gamma_{ijkm}^{\text{den}}$ and $E = 2$. Results show that MTE introduces about 4. 4% and 4. 2% absolute (10. 6% and 10. 9% relative) improvement on recognition rate compared with ML_CI and ML_CD, respectively. The last line in Tab. 1 is from MTE_CD + DM (combination MTE_CD model with duration model, DM), which yields the best recognition results. Tab. 2 shows the details of MTE iterations. It can be seen that 4 and 5 iterations of MTE training obtain the best results.

Tab. 1 Tone classification test results

Tone model	Tone recognition accuracy/%					
	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Average
ML_CI	74. 6	54. 9	43. 2	60. 3	57. 7	58. 68
ML_CD	75. 6	57. 9	48. 3	62. 7	59. 5	61. 37
MTE_CI	75. 5	57. 3	45. 3	71. 1	50. 7	63. 07
MTE_CD	76. 4	60. 7	51. 8	72. 9	49. 0	65. 60
MTE_CD + DM	75. 6	61. 8	53. 1	73. 3	55. 4	66. 51

Tab. 2 Tone recognition accuracy in MTE training %

Tone model	Iteration					
	0(ML)	1	2	3	4	5
MTE_CI	58. 68	61. 13	61. 82	62. 75	62. 93	63. 07
MTE_CD	61. 37	62. 83	64. 01	65. 03	65. 60	65. 28

Tab. 3 gives the results of MTE trained tone models and MPE based model weight training when integrated into the continuous speech recognition task. Recognition is carried out in two passes. The first pass is a normal time-synchronous beam search with the MSR baseline acoustic model and the output of this pass is a word lattice. The second pass is a dynamic programming search within the lattice including an acoustic model and tone models to find the most likely path.

Tab. 3 Large vocabulary continuous speech recognition results

Tasks	System	Tonal syllable accuracy/%	Error reduction/%
Tone model test	Baseline	51. 34	—
	ML_CI	55. 41	8. 4
	ML_CD	56. 53	10. 6
	MTE_CI	57. 46	12. 6
	MTE_CD	58. 39	14. 5
Model weight test	MPE	58. 35	14. 4
	MTE_CI + MPE	61. 12	20. 1
	MTE_CD + MPE	62. 04	23. 0
	MTE_CI + MPE + WT	63. 77	25. 5
	MTE_CD + MPE + WT	64. 62	27. 3

The upper part of Tab. 3 demonstrates the results of integrating different tone models. When the ML_CI is added into the system, the tonal syllable accuracy improves from 51. 34% to 55. 41%. The MTE trained MTE_CI models improve the accuracy to 57. 46%. The MTE_CD model improves the accuracy from 56. 53% to 58. 39%, about 1. 9% absolute (4. 2% relative) better than ML_CD in continuous recognition tasks. From Tab. 1 and Tab. 2, we can see that the MTE trained models are better than the ML trained models in both tone recognition and continuous speech recognition tasks.

The lower part of Tab. 3 shows the results of model weight training (WT). We train η_1 and η_2, and the duration model weight is not considered. Before WT is performed, the model weights are initialized from global weights. The optimal global weights can be selected manually by testing the MPE objective function for all the training utterances. However, the computation may take a lot of time so, for simplicity, this is not done. Instead, the global weights are first selected as $\eta_1 = \eta_2 = 0.5$ for all model combinations. Constant C in Eqs. (14) and (17) is set to $C = E\max_{i=1}^{I}(\kappa\gamma_q^{\text{MPE}} \cdot \eta_{b,i}\log(O_i \mid \xi_i))$, where $E = 300$ is empirically selected by evaluating the MPE objectives of 200 sentences from the training data for speed and convergence. Since the acoustic score and tone score can be pre-computed, the WT training is very efficient (about 0. 01 RxT on a P4 2. 8 GHz CPU). In the training phase, there are a total of 1 806 565 unique model combinations trained. And in the testing phase, those model combinations not trained are given the default values.

The MPE system uses the MPE-trained acoustic model for recognition and the accuracy is 58. 35%. The MTE_CI + MPE and MTE_CD + MPE run recognition with the MPE-trained acoustic model and tone model of MTE_CI and MTE_CD, respectively. MTE_CI + MPE + WT and MTE_CD + MPE + WT is to perform model weight training after the MPE + MTE_CI and MPE + MTE_CD. MTE_CI + MPE + WT and MTE_CI + MPE + WT introduce absolute 2. 7% and 2. 6% improvement, respectively, whereby we can see the effectiveness introduced by MPE based weight training. It is also shown that the MTE_CD + MPE + WT approach leads to the best performance. A relative 27. 3% gain in comparison with the MSR toolbox baseline. Fig. 1 shows the MPE objective function and

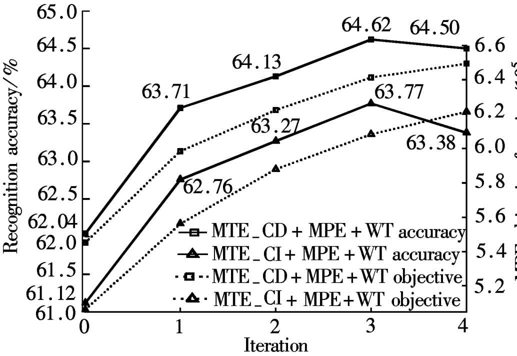

Fig. 1 MPE weight training iterations

recognition accuracy changes. It can be seen from the figure that in the 4th iteration, recognition results drop slightly while the values of the MPE objective function still increase after the 4th iteration, indicating that the WT has reached overtraining.

4 Conclusion

We have evaluated discriminative training of tone models which is referred to as the MTE training. When integrating the discriminatively trained tone models into continuous speech recognition, we consider the use of a discriminative weight optimization based on the MPE criterion. Both methods produced substantial recognition improvements in tone classification and continuous speech recognition tasks. The MPE based model weight training also provides a promising framework for optimal fusion of heterogeneous features or models. It can be easily extended to tone modeling techniques other than HMM such as the SPTM and the decision tree based tone model[2-3] or even the audio-visual speech recognition case.

References

[1] Huang C H, Side F. Pitch tracking and tone features for mandarin speech recognition [C]//*Proceedings of the 25th International Conference on Acoustics, Speech and Signal Processing*. Istanbul, Turkey, 2000: 1523 - 1526.

[2] Cao Yang, Zhang Shuwu, Huang Taiyi, et al. Tone modeling for continuous Mandarin speech recognition [J]. *International Journal of Speech Technology*, 2004, **7**(2/3): 115 - 128.

[3] Wong P F, Siu M H. Decision tree based tone modeling for Chinese speech recognition [C]//*Proceedings of the 29th International Conference on Acoustics, Speech and Signal Processing*. Montreal, Canada, 2004: 905 - 908.

[4] Bahl L R, Brown P F, Souza P, et al. Maximum mutual information estimation of hidden Markov model parameters for speech recognition[C]//*Proceedings of the 11th International Conference on Acoustics, Speech and Signal Processing*. Tokyo, Japan, 1986: 49 - 52.

[5] Juang B H, Chou W, Lee C H. Minimum classification error rate methods for speech recognition [J]. *IEEE Transactions on Speech Audio Processing*, 1997, **5**(2): 266 - 277.

[6] Povey D, Woodland P C. Minimum phone error and I-smoothing for improved discriminative training [C]//*Proceedings of the 27th International Conference on Acoustics, Speech and Signal Processing*. Florida, USA, 2002: 105 - 108.

[7] Povey D. Discriminative training for large vocabulary speech recognition [D]. Peterhouse: Cambridge University, 2004.

[8] Liu Peng, Wang Zuoying. Stream weight training based on MCE for audio-visual LVSCR [J]. *Tsinghua Science and Technology*, 2005, **10**(2): 141 - 144.

[9] Kuo J W, Chen B. Minimum word error based discriminative training of language models[C]//*Proceedings of the 9th European Conference on Speech Communication and Technology*. Lisbon, Portugal, 2005: 1277 - 1280.

[10] Gopalakrishnan P S, Kanevsky D, Nadas A, et al. A generalization of the Baum algorithm to rational objective functions [C]//*Proceedings of the 25th International Conference on Acoustics, Speech and Signal Processing*. Glasgow, Scotland, 1989: 631 - 634.

[11] Chang Eric, Shi Yu, Zhou Jianlai, et al. Speech lab in a box: a Mandarin speech toolbox to jumpstart speech related research [C]//*Proceedings of the 7th European Conference on Speech Communication and Technology*. Aalborg, Denmark, 2001: 2779 - 2782.

[12] Wang Hsin Min, Ho Tai Hsuan, Yang Rung Chiung, et al. Complete language with very large vocabulary but limited training data [J]. *IEEE Transactions on Speech and Audio Processing*, 1997, **5**(2): 196 - 201.

汉语语音识别中区分性声调模型及最优集成方法

黄 浩　朱 杰

(上海交通大学电子工程系,上海 200240)

摘要:提出了 2 种解决汉语语音识别中声调问题的方法:利用区分性方法对基于隐马尔可夫模型(HMM)的声调模型进行训练;提出将区分性训练的声调模型加入大词汇量连续语音识别系统的最优方法,该方法根据最小音子错误的训练准则以及利用扩展 Baum-Welch 算法区分性训练与模型相关的概率权重,对声学模型以及声调模型概率进行加权. 实验结果表明区分性训练的声调模型能够显著地提高连续语音声调识别率以及大词汇量语音识别系统的识别率,同时区分性的模型权重训练能够在区分性声调模型加入连续语音识别系统之后进一步提高系统的识别性能.

关键词:区分性训练;最小音子错误;声调模型;汉语语音识别

中图分类号:TN912