THE JOURNAL OF CHINA UNIVERSITIES OF POSTS AND TELECOMMUNICATIONS
Volume 14, Issue 4, December 2007

HUANG Hao, ZHU Jie

Discriminative tonal feature extraction method in mandarin speech recognition

CLC number TN912 **Document** A **Article ID** 1005-8885 (2007) 04-0126-05

Abstract To utilize the supra-segmental nature of Mandarin tones, this article proposes a feature extraction method for hidden markov model (HMM) based tone modeling. The method uses linear transforms to project F_0 (fundamental frequency) features of neighboring syllables as compensations, and adds them to the original F_0 features of the current syllable. The transforms are discriminatively trained by using an objective function termed as "minimum tone error", which is a smooth approximation of tone recognition accuracy. Experiments show that the new tonal features achieve 3.82% tone recognition rate improvement, compared with the baseline, using maximum likelihood trained HMM on the normal F_0 features. Further experiments show that discriminative HMM training on the new features is 8.78% better than the baseline.

Keywords discriminative training, tone recognition, feature extraction, Mandarin speech recognition

1 Introduction

Tone recognition is an important task for Mandarin speech recognition. Every character in Chinese is pronounced as a base syllable, associated with a tone. There are five tones in Mandarin speech, characterized as high, rising, low, falling, and neutral, which are commonly labeled as tone 1 to tone 5. According to the MSR speech toolbox baseline [1], the recognition rates of tonal and base syllables are 51.3% and 74.8%, respectively, which indicates that 48.3% errors are because of misrecognition of the tones. To improve tone recognition accuracy, considerable efforts have been made to discuss tone modeling. Among these, the HMM-based approach [2] is the mainstream. In addition, other techniques, such as, neural networks (NN) [3], stochastic polynomial trajectory model (SPTM) [4], and decision tree based tone classifier [5] have also been tried.

Although there are basically five lexical tones in Mandarin, tone recognition on continuous speech is much more difficult than the isolated tone recognition task because of the coarticulation effect, that is, tone is not only determined by pitch contour of the current syllable, but also heavily influenced by the behavior of left and right tones. How to characterize tone variation with different contexts has been widely discussed. The context-dependent HMMs [6] were selected by observing the coarticulation effect of neighboring tones were investigated. The decision tree-based clustering method was applied to obtain the optimal context-dependent models [4]. These approaches showed effectiveness in reducing diversity of tone with different contexts; however, they did not really integrate context tone features into either training or recognition.

Of late, the success of the discriminative training methods, such as, maximum mutual information (MMI)[7] and minimum phone error (MPE) [8] has triggered a research focus on discriminative feature optimization [9]. Several feature extraction techniques have been published, including MMI- stereo based piecewise linear compensation for environments (MMI-SPLICE) [10], feature space minimum phone error (fMPE) [11], region dependent linear transformation (RDLT) [9]. They share the similar idea of training linear transform (s) consistent with a certain discriminative criterion on the features.

In this article, the authors focus on the HMM based tone model and show utilization of context F_0 features to improve tone recognition accuracy. They propose discriminative feature extraction (DTFE). The idea is to concatenate the F_0 features of the previous, current, and following syllables to form a long span F_0 feature, and use transforms to project the long span feature as offsets, and then add the offsets to the original F_0 features of the current syllable to create new features. The transforms are trained according to a discriminative training criterion termed as minimum tone error (MTE), which is an approximation of tone recognition accuracy. The proposed method is evaluated on experiments on continuous tone recognition tasks. Results show that maximum likelihood estimation (MLE) on DTFE features give 3.82% (8.86% relative) improvement over the baseline. The authors have also performed discriminative HMM training according to the

Received date: 2007-04-24
HUANG Hao (✉), ZHU Jie
Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
E-mail: haohuang@sjtu.edu.cn

MTE objective function on the DTFE features, which is 8.78% absolute (20.4% relative) better than the baseline. This is also 2.43% absolute better (6.61% relative) than the discriminative HMM training on the normal F_0 features.

2 DTFE

2.1 Tonal feature compensation

Let $\xi = \{\xi_i\}_{i=1}^{5}$ denote the five tones and $\xi_i = \{\xi_{ij}\}_{j=1}^{J_i}$ denote the J_i submodels for tone class ξ_i. Given the normal F_0 frames $\boldsymbol{F}$, the DTFE feature compensation scheme is expressed as:

$$\boldsymbol{y}_{ij} = \boldsymbol{F} + \boldsymbol{w}_{ij}\boldsymbol{x} \tag{1}$$

where $\boldsymbol{y}_{ij}$ is the newly transformed feature frame for model ξ_{ij}. $\boldsymbol{x}$ is a long-span concatenated F_0 features of three successive syllables (previous syllable, current syllable, following syllable) as shown in Fig. 1. As the length of each segment does not always have the same length, the technique of polynomial regression [5] is adopted to make the F_0 feature of each segment have the same dimension of L. In Eq. (1), $\boldsymbol{w}_{ij}$ is the transform matrix associated with model ξ_{ij}:

$$\boldsymbol{w}_{ij} = [w_{ijpq}] = \begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,3L} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,3L} \\ \vdots & \vdots & & \vdots \\ w_{L,1} & w_{L,2} & \cdots & w_{L,3L} \end{pmatrix}_{L\times 3L} \tag{2}$$

where $p = (1, 2, ..., L)$ and $q = (1, 2, ..., 3L)$. Considering the DTFE matrix $\boldsymbol{w}_{ij}$, it can be seen that it will make the DTFE features $\boldsymbol{y}_{ij}$ equal to normal features $\boldsymbol{F}$, if $\boldsymbol{w}_{ij}$ is set at zero in Eq. (2). The aim of DTFE is to train the matrices $\boldsymbol{w}_{ij}$ to obtain new tonal features $\boldsymbol{y}_{ij}$, for better discrimination.

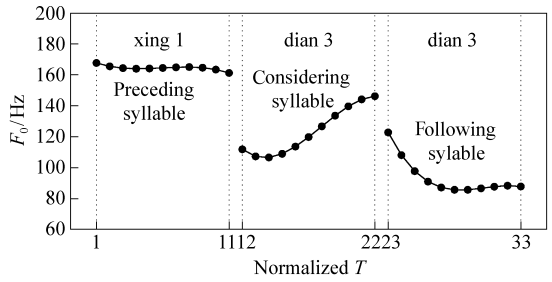

Fig. 1 Time-normalized pitch contours of three successive syllables

2.2 MTE objective function

The transform $\boldsymbol{w}_{ij}$ will be trained discriminatively according to the MTE objective function. Given R tonal syllables and observations O_r for syllable $r\{r = 1, 2, \ldots, R\}$, the MTE objective function is a smooth approximation of tone recognition accuracy and is expressed as a sum of weighted tone accuracy:

$$F_{\mathrm{MTE}} = \sum_{r}^{R}\sum_{i=1,j=1}^{5,J_i} P^k(\xi_{ij} \mid O_r)\delta(\xi_{ij}, \xi_r) \tag{3}$$

where $P(\xi_{ij} \mid O_r)$ is the posterior probability of tone model ξ_{ij} given O_r. $\delta(\xi_{ij}, \xi_r)$ is the accuracy measure and $\delta(\xi_{ij}, \xi_r) = 1$ if ξ_{ij} is the correct tone ξ_r and is zero otherwise. k is a probability scaling factor to reduce the probability dynamic range.

2.3 Updating functions for the DTFE matrix

The transformations are trained using gradient ascent (to maximize F_{MTE}) where the differential of MTE objective function, w.r.t, the weighting matrix, needs to be computed. The authors have used ML model parameter re-estimation after feature updating, as has been done in Ref. [11]. The MTE objective can be written as the function of both the model parameters and the training features: $F_{\mathrm{MTE}}(\mu_{ijsm}, \sigma^2_{ijsm}, y_{ij}(t))$, whereby the differential w.r.t feature can be written as:

$$\frac{\partial F_{\mathrm{MTE}}(\mu_{ijsm}, \sigma^2_{ijsm}, y_{ij}(t))}{\partial y_{ij}(t)} = \frac{\partial F_{\mathrm{MTE}}}{\partial y_{ij}(t)} + \frac{\partial F_{\mathrm{MTE}}}{\mu_{ijsm}}\frac{\partial \mu_{ijsm}}{\partial y_{ij}(t)} + \frac{\partial F_{\mathrm{MTE}}}{\sigma^2_{ijsm}}\frac{\sigma^2_{ijsm}}{\partial y_{ij}(t)} \tag{4}$$

where $y_{ij}(t)$ is the element of $\boldsymbol{y}_{ij}$ at time t. μ_{ijsm} and σ^2_{ijsm} are respectively the mean and variance of mixture m in state s of ξ_{ij} and are also functions of $y_{ij}(t)$, the training features. The first item in Eq. (4) is the derivative of the objective w.r.t feature, by assuming that the model parameters remain fixed:

$$\frac{\partial F_{\mathrm{MTE}}}{\partial y_{ij}(t)} = k\sum_{r}^{R}\sum_{i}^{5}\sum_{j}^{J_i}\gamma_{ij}^{\mathrm{MTE}}\partial\frac{\log P(O_r \mid \xi_{ij})}{\partial y_{ij}(t)} \tag{5}$$

where $\gamma_{ij}^{\mathrm{MTE}} = \gamma_{ij}(\mathrm{Acc}(\xi_{i,j}) - \varepsilon)$. γ_{ij} is the tone model posterior probability from ξ_{ij}. ε is the average tone accuracy for all models. In Eq. (5)

$$\frac{\partial \log P(O_r \mid \xi_{ij})}{\partial y_{ij}(t)} = \sum_{s}^{S}\sum_{m}^{M_s} -\gamma_{ijsm}(t)\frac{\mu_{ijsm} - y_{ij}(t)}{\sigma_{ijsm}} \tag{6}$$

where $\gamma_{ijsm}(t)$ is the model within the Gaussian posterior occupancy of the mixture component m, in state s, in model ξ_{ij}. The second and the third items in Eq. (4) take into account the shift of the model parameters. The differential w.r.t, the mean, and variance in Eq. (4) can be computed by Ref. [11]:

$$\frac{\partial F_{\mathrm{MTE}}}{\partial \mu_{ijsm}} = \frac{k}{\sigma^2_{ijsm}}(\theta^{\mathrm{num}}_{ijsm}(O) - \theta^{\mathrm{den}}_{ijsm}(O) - \mu_{ijsm}(\gamma^{\mathrm{num}}_{ijsm} - \gamma^{\mathrm{den}}_{ijsm})) \tag{7}$$

$$\frac{\partial F_{\mathrm{MTE}}}{\partial \sigma^2_{ijsm}} = \frac{k\gamma^{\mathrm{num}}_{ijsm}}{2}(S^{\mathrm{num}}_{ijsm}\sigma^{-4}_{ijsm} - \sigma^{-2}_{ijsm}) - \frac{k\gamma^{\mathrm{den}}_{ijsm}}{2}(S^{\mathrm{den}}_{ijsm}\sigma^{-4}_{ijsm} - \sigma^{-2}_{ijsm}) \tag{8}$$

where $\gamma_{ijsm}^{\text{num}}$, $\theta_{ijsm}^{\text{num}}(O)$, and S_{ijsm}^{num} are the occupation data, sum-of- data, and scaled variance around the updated mean of the numerator, respectively. $\gamma_{ijsm}^{\text{den}}$, $\theta_{ijsm}^{\text{den}}(O)$, and S_{ijsm}^{den} are the corresponding statistics for the denominator. More details about these statistics can be found in Ref. [12]. The differentials of model parameters w.r.t $\partial y_{ij}(t)$ the feature in Eq. (4) are obtained by taking the derivative of the MLE updating formulae:

$$\frac{\partial \mu_{ijsm}}{\partial y_{ij}(t)} = \frac{\gamma_{ijsm}^{\text{ML}}(t)}{\gamma_{ijsm}^{\text{ML}}} \tag{9}$$

$$\frac{\partial \sigma_{ijsm}^{2}}{\partial \text{y}_{ij}(t)} = \frac{2\gamma_{ijsm}^{\text{ML}}(t)(y_{ij}(t)-\mu_{ijsm})}{\gamma_{ijsm}^{\text{ML}}} \tag{10}$$

where $\gamma_{ijsm}^{\text{ML}}(t)$ is the posterior occupation probability of $y_{ij}(t)$ in MLE for mixture m, in state s. $\gamma_{ijsm}^{\text{ML}}$ is the summation of $\gamma_{ijsm}^{\text{ML}}(t)$ for all the training frames. Then the final differential for training is:

$$\frac{\partial F_{\text{MTE}}}{\partial w_{ijpq}} = \sum_{t=1}^{T} \frac{\partial F_{\text{MTE}}}{\partial y_{ij}(t)} \frac{\partial y_{ij}(t)}{\partial w_{ijpq}} = \sum_{t=1}^{T} x_q \frac{\partial F_{\text{MTE}}}{\partial y_{ij}(t)} \tag{11}$$

where x_q is the qth value of long span feature $\boldsymbol{x}$ and T is the number of total training frames.

2.4 Gradient ascent updating

Gradient ascent updating is expressed as: $w_{ijpq}^{(n+1)} = w_{ijpq}^{(n)} + \eta^{(n)} \partial F_{\text{MTE}} / \partial w_{ijpq}$, where $\eta^{(n)}$ is the learning rate for iteration n. The rule for selecting the learning rate $\eta^{(n)}$ is close to the heuristic approach in Ref. [11], where the differential w.r.t, certain w_{ij}, is accumulated separately for the positive part and negative part at each time t, that is,

$$\nabla_{+} = \sum_{t=1}^{T} \max(\partial F_{\text{MTE}} / \partial w_{ijpq},\ 0) \tag{12}$$

$$\nabla_{-} = \sum_{t=1}^{T} \max(\partial F_{\text{MTE}} / \partial w_{ijpq},\ 0) \tag{13}$$

The learning rate $\eta^{(n)}$ is set to $\eta^{(n)} = E / \nabla_{+} + \nabla_{-}$. Basically, the constant E can be set around $\sigma / 3\mu L$, where by the change of the feature will have the maximum of one standard deviation. The learning rate constant E for speed and convergence can be initialized empirically by evaluating the objective improvement on the cross-validation data (about 1% of the training data in the later experiments). The iterative DTFE training can be summarized as follows:

Step 1 Initialize.

Step 2 Accumulate MTE data for both the numerator and denominator. Note the statistics $\gamma_{ijsm}^{\text{ML}}$ in Eqs. (9)–(10) can be accumulated in this iteration. This is the first forward-backward pass over the training data.

Step 3 Calculating the differential of F_{MTE} w.r.t model parameters using Eqs. (7)–(8).

Step 4 Do the second pass over the training data, accumulating DTFE differentials w.r.t the matrices $\boldsymbol{w}_{ij}$.

Step 5 Update the transformation matrix from $\boldsymbol{w}_{ij}^{(n)}$ to $\boldsymbol{w}_{ij}^{(n+1)}$ using gradient ascent.

Step 6 Do the third pass over the training data, using MLE, updating by using the newly transformed features.

Step 7 Go to step 2, until convergence or maximum number of iterations is reached.

3 MTE based HMM training

Different from DTFE, MTE HMM training maximizes the MTE objective function by updating HMM model parameters. An optimization method using the extended baum-welch (EBW) algorithm [8, 12] is applied. Tone model parameters can be re-estimated iteratively based on parameters of previous iterations by the following formulae:

$$\mu'_{ijkm} = \frac{\left\{\theta_{ijkm}^{\text{num}}(O) - \theta_{ijkm}^{\text{den}}(O)\right\} + D_{ijkm}\mu_{ijkm}}{\left\{\gamma_{ijkm}^{\text{num}} - \gamma_{ijkm}^{\text{den}}\right\} + D_{ijkm}} \tag{14}$$

$$\sigma'^{2}_{ijkm} = \frac{\left\{\theta_{ijkm}^{\text{num}}(O^2) - \theta_{ijkm}^{\text{den}}(O^2)\right\} + D_{ijsm}(\sigma_{ijkm}^{2} + \mu_{ijkm}^{2})}{\left\{\gamma_{ijkm}^{\text{num}} - \gamma_{ijkm}^{\text{den}}\right\} + D_{ijkm}} - \mu_{ijkm}^{2} \tag{15}$$

where μ_{ijkm}, μ'_{ijkm}, σ_{ijkm}^{2}, and $\sigma_{ijkm}^{2\prime}$ are respectively means and variances of current and new estimated values. $\theta_{ijkm}^{\text{num}}(O^2)$ and $\theta_{ijkm}^{\text{den}}(O^2)$ are sum-of-square data for nominator and denominator. D_{ijsm} is a positive smoothing constant. Transition probability and mixture weight will 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 Ref. [8, 12].

4 Experiments and results

The proposed DTFE is evaluated by tone recognition tasks in a large vocabulary of continuous Mandarin speech recognition database. The corpus from microsoft research Asia [1] is used for training. The database contains read speech of about 31.5 hours from 100 male students, for a total of 19 688 utterances and 454 291 tonal syllables. In total there are 100 427 syllables with tone 1, 100 280 syllables with tone 2, 85 055 syllables with tone 3, 154 663 syllables with tone 4, and 26 758 syllables with tone 5. In the testing phase, the MSR [1] test uses an additional 0.74 hour, 500 utterances (9 570 syllables in total) from another 25 male speakers. The speech waveforms are sampled at 16 bit and 16 kHz. To obtain the voiced portion of a syllable, voiced/unvoiced detection

algorithm proposed in Ref. [13] can be used. In this article, the authors have achieved this by using HMM forced alignment.

In addition to the normal F_0 features or DTFE features, normalized log energy and its first and second derivatives (E, ΔE, and $\Delta\Delta E$), first derivative of $F_0(\Delta F_0)$ are also used. For tone modeling, a set of 175 context dependent tone models are used; that is, 5^2 (at the beginning of the sentence) $+5^3$ (in the middle of a sentence), and 5^2 (at the end of the sentence). Each tone HMM has three emitting states with eight Gaussians per state. Scaling factor k is selected as $k = 17$ in both DTFE and MTE HMM training. Smoothing constant D in Eqs. (14)–(15) is typically selected as $D = 2\gamma_{ijsm}$.

Tables 1–3 shows the comparison of tone recognition results using different features (normal F_0 feature or DTFE feature) with different HMM training methods (MLE HMM training or MTE discriminative training).

Table 1 Tone recognition accuracy of DTFE

	Tone recognition accuracy/%					
	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Average
Baseline	68.8	54.5	41.5	58.1	64.5	56.9
DTFE	69.9	57.4	52.4	63.1	54.1	60.7

Table 1 gives the results of baseline and DTFE. Baseline uses model trained with MLE on the normal F_0 and the accuracy is 56.9%. DTFE is the result of using HMM trained with MLE on DTFE features, which is 3.82% better than the baseline. Note that the recognition rate of tone 3 in DTFE rises from 41.5% to 52.4%, indicating that tone 3 recognition is most significantly dependent on its context.

Table 2 Results of MTE HMM training on normal F_0 features

	Tone recognition accuracy/%					
τ	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Average
0	74.9	61.0	44.2	75.3	25.6	63.2
50	76.5	60.4	45.1	72.3	39.8	63.5
100	76.3	59.6	46.0	70.8	43.9	63.2
150	75.9	59.5	47.0	69.7	45.9	63.1
200	75.6	58.6	47.3	68.9	47.1	62.7

Table 2 shows the results of MTE HMM training on the normal F_0 features. I-smoothing [8] with different constant τ, which is evaluated to avoid individual performance degradation, as there is far less training data for tone 5. It is shown through the overall results that using different smoothing constant τ close to each other, the results of tone 5 do not degrade too much when τ increases. It is also shown the MTE HMM training improves the accuracy to 63.2% when no smoothing is applied ($\tau = 0$). The best result is 63.5% ($\tau = 50$) (about 6.57% absolute and 15.2% relative improvement over the baseline)

Table 3 demonstrates the results of DTFE + MTE, that is, to perform MTE HMM training on DTFE features. As shown, DTFE + MTE is significantly better than MTE HMM training on the normal F_0 features in Table 1. The best results is 65.7%, about 2.43% better than the MTE model training on normal F_0 features and 8.78% better than the baseline. From the above statistics it can be seen that the DTFE features outperform the normal F_0 features under various HMM training methods.

Table 3 Results of MTE HMM training on DTFE features

	Tone Recognition Accuracy/%					
τ	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Average
0	76.5	59.0	48.7	77.3	40.2	65.7
50	76.7	60.1	51.7	73.8	41.9	65.3
100	76.5	59.1	51.8	74.2	45.1	65.5
150	76.3	59.8	50.7	72.1	50.5	65.1
200	76.4	60.0	51.5	72.0	51.3	65.3

In DTFE, the objective improvement of F_{MTE} (expressed as F_{MTE} divided by the number of training syllables) is 4.61% (from 0.417 to 0.463), obtaining 3.82% improvement in accuracy. In the MTE HMM training, the objective improvement is 19.7% (from 0.417 to 0.614), obtaining only 6.57% improvement in accuracy. This indicates that DTFE is more robust to overtraining than MTE HMM training. As for the training speed, DTFE converges in about 12–14 iterations. Learning rate constant E for DTFE is increased from 0.003 to 0.013, step by step, to avoid fast convergence. In MTE HMM training, six to seven iterations can obtain the best performance.

Figure 2 illustrates the original F_0, projected offsets, and compensated F_0 for $T_1 - T_3 + T_1$, that is, F_0 feature of tone 3 with a left and a right context of tone 1. It is obvious that the original F_0 contour is "flattened" by the projected offset, whereby one thinks that the physical meaning of DTFE is trying to eliminate the coarticulating effect adaptively according to the tone contexts. It can also be seen that the current syllable is more influenced by its left context, which is consistent with the conclusion drawn in Ref. [5].

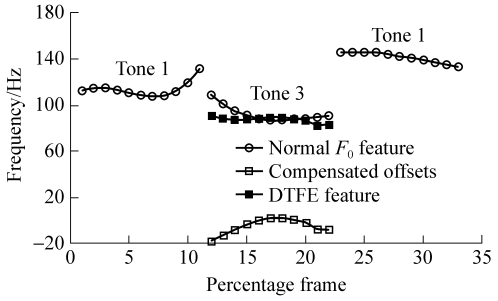

Fig. 2 DTFE projected offset and new feature

5 Conclusions

In this article, the authors have proposed DTFE. The method is to discriminatively train matrices to project long span F_0 features and add the projected offsets to the F_0 features of the current syllable. Experiments of tone recognition on

continuous Mandarin speech show that the DTFE derived features have significantly outperformed the normal F_0 features. And further improvement has been yielded by the overall training of both the frond-end feature extraction and the back-end model parameters.

References

1. Chang E, Shi Yu, Zhou Jian-lai, et al. Speech lab in a box: a Mandarin speech toolbox to jumpstart speech related research. Proceedings of the 7th European Conference on Speech Communication and Technology, Sep 3−7, 2001, Aalborg, Denmark. 2001: 2779−2782
2. Huang H C H, Seide F. Pitch tracking and tone features for mandarin speech recognition. Proceedings of International Conference on Acoustics, Speech and Signal Processing: Vol 3, Jun 5−9, 2000: Istanbul, Turkey. Piscataway, NJ, USA: IEEE, 2000: 1523−1526
3. Thubthong N, Kijsirikul B. Tone recognition of continuous Thai speech under tonal assimilation and declination effects using half-tone model. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2001. 9(6): 815−825
4. Cao Yang, Zhang Shu-wu, Huang Tai-yi, et al. Tone modeling for continuous Mandarin speech recognition. International Journal of Speech Technology, 2004, 7(2−3): 115−128
5. Wong P F, Siu M H. Decision tree based tone modeling for Chinese speech recognition. Proceedings of International Conference on Acoustics, Speech and Signal Processing: Vol 1, May 17−21, 2004: Montreal, Canada. Piscataway, NJ, USA: IEEE, 2004: 905−908
6. Wang H M, Ho T H, Yang R C, et al. Complete language with very large vocabulary but limited training data. IEEE Transactions on Speech and Audio Processing, 1997, 5(2): 196−201
7. Bahl L R, Brown P F, Souza P, et al. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. Proceedings of International Conference on Acoustics, Speech and Signal Processing, Vol.1, Apr. 18−22, 1986, Tokyo, Japan. Piscataway, NJ, USA: IEEE, 1986: 49−52
8. Povey D, Woodland P C. Minimum phone error and I-smoothing for improved discriminative training. Proceedings of International Conference on Acoustics Speech and Signal Processing, Vol.1, May. 13−17, 2002, Orlando, FL, USA. Piscataway, NJ, USA: IEEE, 2002: 105−108
9. Zhang B, Matsoukas S, Schwartz R. Discriminatively trained region dependent feature transforms for speech recognition. Proceedings of International Conference on Acoustics, Speech and Signal Processing: Vol.1, May 14−19, 2006, Toulouse, France. Piscataway, NJ, USA: IEEE, 2006: 313−316
10. Droppo J, Deng L, Acero A. Evaluation of the SPLICE algorithm on the Aurora2 database. Proceedings of Eurpean Conference on Speech Communication Technology: Vol. 1, Sep 3−7, 2001, Aalborg, Denmark. 2001: 217−220
11. Povey D, Kingsbury B, Mangu L, et al. fMPE: discriminatively trained features for speech recognition. Proceedings of International Conference on Acoustics, Speech and Signal Processing: Vol. 1, Mar 18−23, 2005, Philadelphia, PA, USA. Piscataway, NJ, USA: IEEE, 2005: USA, 961−964
12. Povey D. Discriminative Training for Large Vocabulary Speech Recognition. Ph D, Cambridge, UK: Cambridge University, 2004
13. Ying Na, Zhao Xiao-hui, Dong Jing. Unvoiced/voiced classification and voiced harmonic parameters estimation using the third-order statistics. The Journal of China Universities of Posts and Telecommunications, 2007, 14(1): 85−89

Biographies: HUANG Hao, Ph. D. Candidate of Shanghai Jiaotong University, interested in the research on speech recognition and speech signal processing.

ZHU Jie, Ph. D., professor, Shanghai Jiao Tong University, interested in the research on speech recognition and human-machine interaction.