Using vector Taylor series with noise clustering for speech recognition in non-stationary noisy environments①

Zhao Xianyu (赵贤宇)②, Ou Zhijian, Wang Zuoying
(Department of Electronic Engineering, Tsinghua University, Beijing 100084, P.R.China)

Abstract

The performance of automatic speech recognizer degrades seriously when there are mismatches between the training and testing conditions. Vector Taylor Series (VTS) approach has been used to compensate mismatches caused by additive noise and convolutive channel distortion in the cepstral domain. In this paper, the conventional VTS is extended by incorporating noise clustering into its EM iteration procedure, improving its compensation effectiveness under non-stationary noisy environments. Recognition experiments under babble and exhibition noisy environments demonstrate that the new algorithm achieves 35% average error rate reduction compared with the conventional VTS.

Key words: speech recognition, robustness, model adaptation, clustering

0 Introduction

The problem of environmental robustness is one of the most important issues in speech recognition. The performance of current automatic speech recognizer degrades seriously when there are mismatches between the training and testing conditions, e.g. due to the additive noise and channel distortion of the environment. One solution to the problem of mismatch is to adapt the acoustic model trained from clean speech to match the current working environment. However, traditional speaker adaptation methods, such as MLLR and MAP, are not applicable because they require large amounts of adaptation data[1, 2]. Parallel Model Compensation (PMC)[3] and Vector Taylor's Series (VTS)[4-6] are now two main methods for acoustic model compensation and adaptation. Because of the nonlinear nature of the corruption model for cepstral parameters under the condition of additive noise and channel distortion, some approximations need to be made to facilitate efficient computation. PMC uses lognormal approximation for this nonlinear model. VTS approximates the non-linear model with its first-order vector Taylor's series expansion, and transforms it into a linear one. In Ref. [6], Monte Carlo experiments were tried to compare VTS and PMC. The analysis results showed that the VTS approximation was more accurate than the lognormal approximation used in PMC. Furthermore, in VTS the maximum likelihood estimates of environmental parameters can be obtained through an Expectation-Maximization (EM) procedure under appropriate Gaussian assumptions. Recognition experiments in Ref. [4 – 6] showed the effectiveness of VTS to improve the environmental robustness of speech recognizer.

In real applications, at first voice activity detection (VAD) algorithm is used to segment a continuous speech stream into several utterances. Then, for each utterance, the VTS algorithm is applied to do environmental parameter estimation and model adaptation. Environmental statistics obtained from the silence segment before each utterance provide the starting point for EM iteration. Under conventional VTS settings, within each utterance, the environment is assumed to be stationary. And only one environmental model is constructed for the environment. However, in time-varying noisy environments, the statistical nature of noise is in fact inhomogeneous across different regions of one utterance. In this paper, unsupervised clustering technique is incorporated into VTS to further improve the accuracy of environmental model estimation and acoustic model compensation within one utterance. Noisy speech frames are not treated as being corrupted by a single stationary noise source. Instead, they are clustered into several environmental classes according to their a posteriori distribution. Separate environmental models were constructed respectively for each class.

The paper is organized as follows. In section 1, the conventional VTS model compensation and parameter estimation using EM are described. In section 2, unsupervised noise clustering is discussed to refine the environmental model. The speech recognition experiments and results are presented in section 3. Finally, conclusions are drawn in section 4.

① Supported by the High Technology Research and Development Programme of China (No.2001AA114071).
② To whom correspondence should be addressed. E-mail: zhaoxy00@mails.tinghua.edu.cn
Received on Aug. 19, 2004

1 VTS model compensation and parameter estimation

The speech corruption model used in this paper is depicted in Fig. 1.

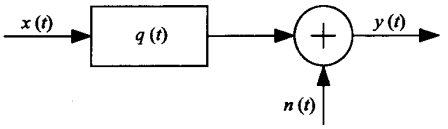

Fig. 1 Speech corruption model in noisy environments

The relationship between clean speech x, corrupted speech y, additive noise n and channel distortion q is

$$y(t) = x(t) \otimes q(t) + n(t) \tag{1}$$

where $\otimes$ stands for convolution. In the cepstral domain, such corruption model can be represented as[5]

$$y_c = x_c + q_c + C \cdot \log(1 + exp(C^{-1} \cdot (n_c - x_c - q_c))) \tag{2}$$

where y_c, x_c, n_c and q_c correspond to the cepstral parameters of y, x, n and q respectively, C is the DCT transformation matrix. In Ref. [5], Vector Taylor's Series (VTS) is proposed to approximate this nonlinear corruption model with its first-order vector Taylor series expansion around (μ_x, μ_n, μ_q),

$$y_c \approx A \cdot (x_c - \mu_x) + B \cdot (n_c - \mu_n) + Q \cdot (q_c - \mu_q) + g(\mu_x, \mu_n, \mu_q) \tag{3}$$

where A, B and Q are the Jacobian of (3) with respect to x_c, n_c and q_c evaluated at (μ_x, μ_n, μ_q), i.e.

$$\begin{cases} A = \dfrac{\partial y_c}{\partial x_c}\Big|_{(\mu_x,\mu_n,\mu_q)} \\ B = \dfrac{\partial y_c}{\partial n_c}\Big|_{(\mu_x,\mu_n,\mu_q)} \\ Q = \dfrac{\partial y_c}{\partial q_c}\Big|_{(\mu_x,\mu_n,\mu_q)} \\ g(\mu_x,\mu_n,\mu_q) = \mu_x + \mu_q + C \cdot \log(1 + \exp(C^{-1} \cdot (\mu_n - \mu_x - \mu_q))) \end{cases} \tag{4}$$

It is usually assumed that x_c and n_c are independent of each other, and both are Gaussian, i.e.

$$x_c \sim N(\mu_x, \Sigma_x) \tag{5}$$

$$n_c \sim N(\mu_n, \Sigma_n) \tag{6}$$

Channel distortion q_c is assumed to be an unknown constant vector whose current estimate is μ_q. Under these assumptions and through the linear approximation (3), the model parameters trained from clean speech, $\Lambda_x = \{\mu_x, \Sigma_x\}$, can be compensated effectively with the estimated environmental characteristics, $\Lambda_n = \{\mu_n, \Sigma_n, \mu_q\}$, to obtain an adapted model, $\Lambda_y = \{\mu_y, \Sigma_y\}$, that matches the current working condition,

$$\begin{cases} \mu_y = g(\mu_x, \mu_n, \mu_q) \\ \Sigma_y = A \cdot \Sigma_x \cdot A^T + B \cdot \Sigma_n \cdot B^T \end{cases} \tag{7}$$

The environmental parameters are estimated through EM algorithm. The parameter estimates of the $(t+1)$-th iteration, $\Lambda_n^{(t+1)} \triangleq \{\mu_n^{t+1}, \Sigma_n^{(t+1)}, \mu_q^{(t+1)}\}$, is found through the maximization of an auxiliary function, Q, which is defined as

$$Q(\Lambda_n, \Lambda_n^{(t)}) = E[\log p(x_c, n_c \mid \{\Lambda_x, \Lambda_n\}) \mid y_c, \{\Lambda_x, \Lambda_n^{(t)}\}] \tag{8}$$

and

$$\Lambda_n^{(t+1)} = \arg\max_{\Lambda_n} Q(\Lambda_n, \Lambda_n^{(t)}) \tag{9}$$

where $\Lambda_n^{(t)}$ corresponds to the estimates of the t-th iteration. For acoustic models using Gaussian Mixture Model (GMM), the solution of Eq. (8) can be written as

$$\mu_n^{(t+1)} = \frac{\sum_{i=1}^{N}\sum_{s=1}^{S}\sum_{k=1}^{K_s} \gamma_i(\omega_{s,k}) \mu_n(y_i, \omega_s, k)}{\sum_{i=1}^{N}\sum_{s=1}^{S}\sum_{k=1}^{K_s} \gamma_i(\omega_{s,k})} \tag{10}$$

$$\Sigma_n^{(t+1)} = \frac{\sum_{i=1}^{N}\sum_{s=1}^{S}\sum_{k=1}^{K_s} \gamma_i(\omega_{s,k}) C_n(y_i, \omega_s, k)}{\sum_{i=1}^{N}\sum_{s=1}^{S}\sum_{k=1}^{K_s} \gamma_i(\omega_{s,k})} - \mu_n^{(t+1)} \cdot \mu_n^{(t+1)^T} \tag{11}$$

$$\mu_q^{(t+1)} = \mu_q^{(t)} + [\sum_{i=1}^{N}\sum_{s=1}^{S}\sum_{k=1}^{K_s} \gamma_i(\omega_{s,k}) B A^{-1} \Sigma_{s,k}^{-1} A^{-1} B]^{-1} \cdot [\sum_{i=1}^{N}\sum_{s=1}^{S}\sum_{k=1}^{K_s} \gamma_i(\omega_{s,k}) B A^{-1} \Sigma_{s,k}^{-1} A^{-1} q(y_i, w_{s,k})] \tag{12}$$

where $\{y_i; i = 1, 2, \cdots, N\}$ is the input observation feature sequence, S is the number of states in the acoustic models, K_s is the number of Gaussian mixture components of state s, $\omega_{s,k}$ stands for the k-th Gaussian mixture component of state s, $\mu_n(y_i, w_{s,k})$ and $C_n(y_i, \omega_{s,k})$ are respectively the mean and correlation matrix of the a posteriori distribution of n_c

$$\begin{cases} \mu_n(y_i, \omega_{s,k}) = E[n_c \mid y_i, \omega_{s,k}, \Lambda_x, \Lambda_n^{(t)}] \\ C_n(y_i, \omega_{s,k}) = E[n_c n_c^T \mid y_i, \omega_{s,k}, \Lambda_x, \Lambda_n^{(t)}] \end{cases} \tag{13}$$

They are calculated through

$$\begin{cases} \mu_n(y_i, \omega_{s,k}) = \tilde{\Sigma}_n(\omega_{s,k}) \cdot (\Sigma_n + \tilde{\Sigma}_n(\omega_{s,k}))^{-1} \cdot \mu_n + \Sigma_n \cdot (\Sigma_n + \tilde{\Sigma}_n(\omega_{s,k}))^{-1} \cdot \tilde{\mu}_n(i, \omega_{s,k}) \\ \mu_n(y_i, \omega_{s,k}) = \tilde{\Sigma}_n(\omega_{s,k}) \cdot (\Sigma_n + \tilde{\Sigma}_n(\omega_{s,k}))^{-1} \cdot \Sigma_n + \mu_n(y_i, \omega_{s,k}) \cdot \mu_n(y_i, \omega_{s,k})^T \end{cases} \tag{14}$$

where

$$\begin{cases} \tilde{\mu}_n(i,\omega_{s,k}) = \mu_n^{(t)} + (I-A)^{-1}(y_i - \\ \qquad g(\mu_{s,k},\mu_n^{(t)},\mu_q^{(t)})) \\ \tilde{\Sigma}_n(\omega_{s,k}) = (I-A)^{-1}A \cdot \Sigma_{s,k} \cdot A(I-A)^{-1} \end{cases} \tag{15}$$

$q(y_i,\omega_{s,k})$ is given as

$$q(y_i,\omega_{s,k}) = y_i - g(\mu_x,\mu_n,\mu_q) - \mu_{s,k} - (I-A)(\mu_n(y_i,\omega_{s,k}) - \mu_n^{(t)}) \tag{16}$$

$\gamma_i(\omega_{s,k}) = p(\omega_{s,k} \mid y_i, \Lambda_n^{(t)})$ is the a posteriori probability of mixture component $\omega_{s,k}$ given observation y_i.

The VTS iterative parameter estimation and recognition procedure are summarized as follows, also depicted in Fig.2.

1. Use the currently estimated parameters to compensate HMM parameters as in Eq.(7);

2. Find the maximum likelihood word sequence $W = \{W_1, \cdots, W_L\}$ embedded in observation sequence Y through Viterbi decoding;

3. Decide if the likelihood increase of $P(Y, W \mid \Lambda_x, \Lambda_n^{(t+1)})$ compared with $P(Y, W \mid \Lambda_x, \Lambda_n^{(t)})$ is significant; if so, using Eqs.(10)-(12) to update parameter estimations and go to step 1; otherwise go to step 4;

4. Stop iteration procedure and output recognition results;

In this procedure, the maximum likelihood estimate of word sequence, W and environmental parameters, Λ_n are jointly optimized[7].

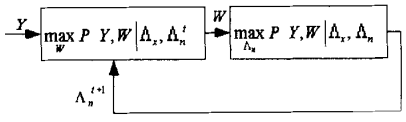

Fig.2 VTS parameter estimation and recognition

2 VTS with noise clustering

In the above development, it is assumed that the noisy environment is stationary during one utterance. Thus all noisy frames are used to estimate the single environment model for additive noise and channel distortion. In order to accommodate the non-stationary nature of the time-varying environment within one utterance, unsupervised clustering technique is combined with VTS in this paper. Noisy speech frames are clustered according to their a posteriori distribution. Frames of similar statistical properties are clustered together. And separate environmental models are constructed and estimated for each class respectively. When doing recognition, different sets of HMM models compensated for different classes are applied to frames belonging to different environmental classes to do Viterbi match.

For each noisy frame, the a posteriori distribution of the environmental noise given observation y_i is assumed to be Gaussian. Its mean, $\mu_{n,i}$, and correlation matrix, $\Sigma_{n,i}$, are calculated by

$$\mu_{n,i} = E[n_c \mid y_i, \Lambda_x, \Lambda_n] = \sum_{s=1}^{S}\sum_{k=1}^{K_s} \gamma_i(\omega_{s,k})\mu_n(y_i,\omega_{s,k}) \tag{17}$$

$$\Sigma_{n,i} = E[n_c n_c^T \mid y_i, \Lambda_x, \Lambda_n] - \mu_{n,i} \cdot \mu_{n,i}^T = \sum_{s=1}^{S}\sum_{k=1}^{K_s} \gamma_i(\omega_{s,k}) C_n(y_i,\omega_{s,k}) - \mu_{n,i} \cdot \mu_{n,i}^T \tag{18}$$

The distance measure used in the clustering procedure is Kullback-Leibler (K-L) measure[8]. For two Gaussian distribution, $p_1(x) \sim N(\mu_1, \Sigma_1)$ and $p_2(x) \sim N(\mu_2, \Sigma_2)$, their K-L distance is defined to be

$$D_{KL}(p_1,p_2) = \int p_1 \cdot \log\frac{p_1}{p_2}dx + \int p_2 \cdot \log\frac{p_2}{p_1}dx = -D + \frac{1}{2}tr(\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1) + (\mu_1-\mu_2)^T(\Sigma_1^{-1}+\Sigma_2^{-1})(\mu_1-\mu_2) \tag{19}$$

where D is the dimension of variable x.

The clustering algorithm uses a hierarchical, incremental refinement process[9]. The VTS algorithm incorporating noise clustering proceeds as follows:

1. Initialization: all noisy frames are assigned to one class.

2. EM algorithm is applied to each class to iteratively estimate their corresponding environmental model until convergence (in our experiments, we judge convergence by the criterion that the relative increase of log-likelihood between adjacent EM iterations is less than 1%). For each class c, use the following formulae (similar to Eqs. (10)-(12)) to iteratively update its parameters, $\Lambda_{n|c}^{(t+1)} \triangleq (\mu_{n,|c|}^{(t+1)}, \Sigma_{n,|c|}^{(t+1)}, \mu_{q,|c|}^{(t+1)})$,

$$\mu_{n,|c|}^{(t+1)} = \frac{\sum_{i\in c}\sum_{s=1}^{S}\sum_{k=1}^{K_s}\gamma_i(\omega_{s,k})\mu_n(y_i,\omega_{s,k})}{\sum_{i\in c}\sum_{s=1}^{S}\sum_{k=1}^{K_s}\gamma_i(\omega_{s,k})} \tag{20}$$

$$\Sigma_{n,|c|}^{(t+1)} = \frac{\sum_{i\in c}\sum_{s=1}^{S}\sum_{k=1}^{K_s}\gamma_i(\omega_{s,k})C_n(y_i,\omega_{s,k})}{\sum_{i\in c}\sum_{s=1}^{S}\sum_{k=1}^{K_s}\gamma_i(\omega_{s,k})} - \mu_{n,|c|}^{(t+1)} \cdot \mu_{n,|c|}^{(t+1)^T} \tag{21}$$

$$\mu_{q,|c|}^{(t+1)} = \mu_{q,|c|}^{(t)} + [\sum_{i\in c}\sum_{s=1}^{S}\sum_{k=1}^{K_s}\gamma_i(\omega_{s,k})BA^{-1}\Sigma_{s,k}^{-1}A^{-1}B]^{-1} \cdot [\sum_{i\in c}\sum_{s=1}^{S}\sum_{k=1}^{K_s}\gamma_i(\omega_{s,k})BA^{-1}\Sigma_{s,k}^{-1}A^{-1}q(y_i,w_{s,k})] \tag{22}$$

where $i \in c$ means the i-th frame y_i is assigned to environmental class c, other parameters are defined the same as in Eqs. (10)-(16).

3. For each class, a distance matrix, D_c, is constructed. Each element, $D_c(i,j)$, of the distance matrix is the K-L distance of the a posteriori noise distributions for the i-th frame and j-th frame assigned to class c, i.e.

$$D_c(i,j) = -D + \frac{1}{2} tr(\Sigma_{n,i}^{-1} \cdot \Sigma_{n,j} + \Sigma_{n,j}^{-1} \cdot \Sigma_{n,i}) + (\mu_{n,i} - \mu_{n,j})^T (\Sigma_{n,i}^{-1} + \Sigma_{n,j}^{-1})(\mu_{n,i} - \mu_{n,j}) \quad (23)$$

4. Define the dispersion measure, S_c, of class to be

$$S_c = \max_{i,j \in c} D_c(i,j) \quad (24)$$

5. Mark all classes as being able to be split.

6. Decide if there are classes being able to be split. If not, go to step 11.

7. Among all classes being able to be split, choose the class $c_{\max}$ with the largest dispersion measure to do split, i.e.

$$c_{\max} = \arg\max_{c \in C} S_c \quad (25)$$

8. Split class $c_{\max}$ into two sub-classes, $c_{\max,1}$ and $c_{\max,2}$. Choose the i_0-th and j_0-th frame of $c_{\max}$, which have the largest K-L distance between all frames of class $c_{\max}$, as the centers of the two new sub-classes, i.e.

$$(i_0, j_0) = \arg\max_{i,j \in c_{\max}} D_{c_{\max}}(i,j) \quad (26)$$

9. Other frames of class $c_{\max}$ are re-assigned to $c_{\max,1}$ and $c_{\max,2}$ according to their distances to the two new centers,

$$\begin{cases} i \in c_{\max,1}, & if\ D_{c_{\max}}(i,i_0) < D_{c_{\max}}(i,j_0) \\ i \in c_{\max,2}, & otherwise \end{cases} \quad (27)$$

10. Decide if the two new created classes have enough frames (above a threshold, in our experiments this threshold is set to be 20) for reliable parameter estimation. If so, go to step 2 to use EM algorithm to re-estimate the environmental parameters for each class. If not, mark current class $c_{\max}$ as not being able to be split, go to step 6 to choose another class.

11. Output recognition results.

3 Experiments and results

A baseline speech recognition system is constructed to recognize continuous digits. The continuous digits corpus consists of 2937 sentences from 20 speakers for the training purpose, while 220 sentences from other two speakers (not appeared in the training set) are used as the test set. Each sentence contains 4, 5 or 8 digits. Speech is digitized with 8 kHz sampling and 16 bit precision. This baseline system consists of 10 HMM's to model Chinese digits from zero to nine, each HMM has six emitting states and 14-mixture GMM is used for each state. An additional state is used to model the silence. The feature vector is formed by 14 MFCC's, energy plus their first and second order differentials. The speech frame length is 20ms and there exists 10ms overlap between adjacent frames.

3.1 Noise clustering experiments

In the first set of experiments, we analyze the effects of the unsupervised noise clustering technique on the VTS algorithm. Figs.3 and 4 show two segments of noise. The noise segment shown in Fig.3 represents the stationary Gaussian white noise. In Fig.4, the noise segment is not stationary, which consists of two different types of noise. The first half is a kind of destroy engine noise, and the second half is machine gun noise. A speech segment is shown in Fig.5. In one experiment, this speech segment is corrupted by the stationary Gaussian white noise segment. Then in Fig.6, we show the log-likelihood of the

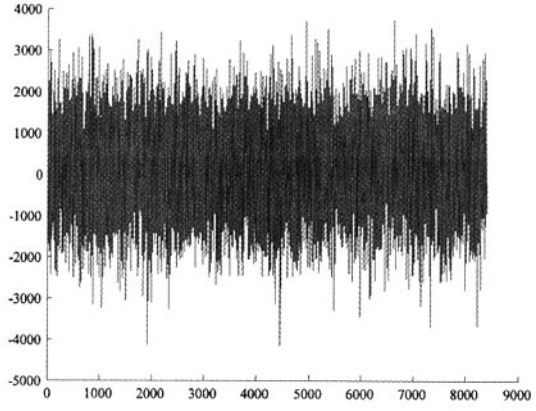

Fig.3 Stationary noise

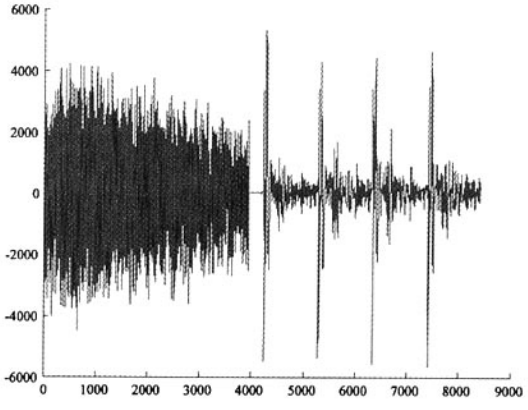

Fig.4 Non-stationary noise

noisy utterance during the EM iteration procedure of the conventional VTS. From this figure, it can be seen that the log-likelihood increases significantly after several iterations with the EM algorithm to identify the statistical characteristics of the corruption noise.

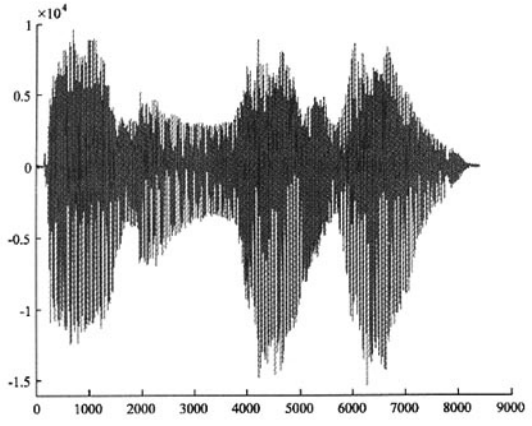

Fig.5 Speech segment

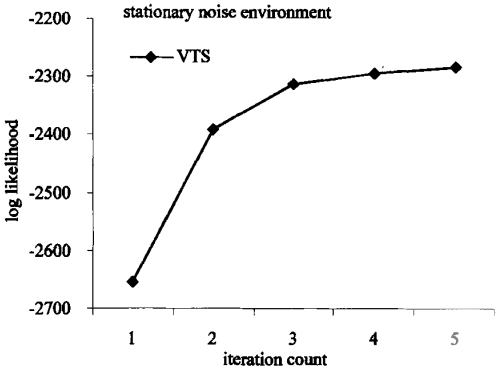

Fig.6 EM iteration for stationary noise

A similar experiment is conducted with the speech segment corrupted by the non-stationary noise as shown in Fig.4. The log-likelihood during EM iteration is shown in Fig.7. But in this case, it can be seen that the log-likelihood converges quickly to a plateau. Thus the models in the conventional VTS are not compensated effectively under non-stationary environments like the case shown in Fig.4. This indicates that the use of only one noise model is not sufficient in this case. After the conventional VTS is extended by incorporating noise clustering into its EM iteration procedure, Fig.7 also shows the log-likelihood during this new algorithm's EM iteration procedure. It starts with one noise model. The first three steps are identical to the conventional VTS. Later, unsupervised noise clustering is carried out to split the original noise model into two sub-models. With the addition of another class of noise model, it is clear form Fig.7 that the log-likelihood is increased significantly in the subsequent iteration steps. In Fig.8 we show how the noisy frames are clustered into the two classes. It can be seen that this clustering approximately reflects the non-stationary nature of the noise segment in Fig.4 (one half destroy engine and one half machine gun). It is shown that the proposed unsupervised clustering algorithm can identify the underlying non-stationary nature successfully, and with this clustering and the split of noise model the effectiveness of model compensation can be improved.

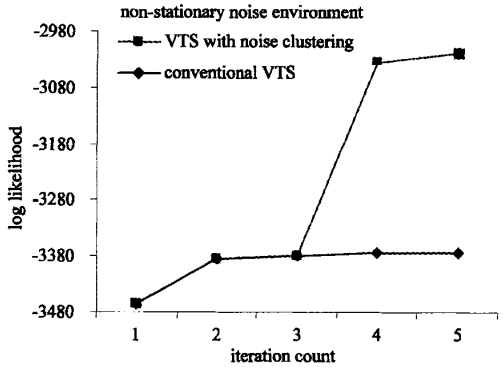

Fig.7 EM iteration for non-stationary noise

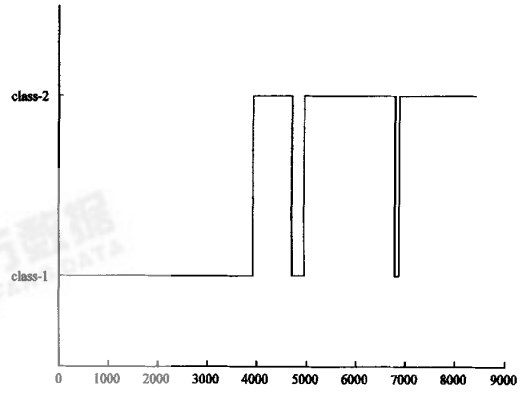

Fig.8 Class information after clustering

In the VTS algorithm, the computation complexity is dominated by the calculation of various a posteriori probabilities. So the incorporation of noise clustering does not increase the computation complexity of VTS significantly.

3.2 Speech recognition experiments

In the second set of experiments, the speech recognition performance of the new algorithm is analyzed and compared with some existing algorithms. These recognition experiments are carried out under babble and exhibition noise environments. Both types of noise were extract-

ed from AURORA noise database provided by European Telecommunications Standard Institute (ETSI)[10]. In these experiments, only the effect of additive noise was considered, while leaving the effect of channel distortion for future investigation. The results are presented as word error rates (WER). WER results under different noise types and Signal-Noise-Ratio (SNR) settings are summarized in Table 1, in which VTS-CLS stands for VTS algorithm with noise clustering as proposed in this paper.

Table 1 Speech recognition WER results under noisy environments

SNR (dB)	Word Error Rate (%)					
	Babble			Exhibition		
	Baseline	VTS	VTS-CLS	Baseline	VTS	VTS-CLS
20	3	2.04	1.72	5.49	2.86	2.34
15	5.17	2.34	1.76	11.08	3.87	2.28
10	12.39	4.41	2.94	29.77	6.34	3.11
5	30.17	12.49	7.48	60.13	14.42	7.39
0	59.47	32.65	24.81	83.26	39.51	18.54

From these results, it can be seen that when SNR drops, the performance of the baseline system degrades seriously. After VTS algorithm is used to adapt the acoustic models, performance improvements are obtained against the baseline. Through using noise clustering in VTS, the system's environmental robustness is further improved in these non-stationary environments. The error rate reduction compared with the conventional VTS is 35%, averaging over different noise types and SNR settings.

4 Conclusion

In this paper, we investigate the use of unsupervised noise clustering to improve the environmental robustness of VTS algorithm in time-varying non-stationary environments. Noisy speech frames in one utterance are clustered into several environmental classes based on their a posteriori distribution and Kullback-Leibler distance measure. Instead of using only one environmental model to compensate/adapt acoustic model as in the conventional VTS, refined environmental models are constructed respectively for each class and EM algorithm is used to iteratively estimate parameters of these models. Recognition experiments under AURORA babble and exhibition environments show that the proposed method achieves better results compared with the conventional VTS algorithm. The system robustness is improved significantly under non-stationary environments with the new method.

References

[1] Leggetter C J, Woodland P C. Maximum likelihood linear regression for speaker adaptation of continuous density Markov models. *Computer Speech and Language*, 1995, 9(2): 171-185

[2] Lee C H, Lin C H, Juang B H. A study on speaker adaptation of parameters of continuous density hidden Markov models. *IEEE Transaction on Acoustics Speech Signal Processing*, 1991, 39(4): 806-814

[3] Gales M J F, Young S J. Robust continuous speech recognition using parallel model combination. *IEEE Transaction on Speech and Audio Processing*, 1996, 4(5): 352-359

[4] Moreno P J, Raj B. A vector Taylor series approach for environmental independent speech recognition. ICASSP'96, 1996: 733-736

[5] Kim D Y, Un C K, Kim N S. Speech recognition in noisy environments using first-order vector Taylor series. *Speech Communication*, 1998, 24(1): 49-49

[6] Acero A, Deng L, Kristjansson T, et al. HMM adaptation using vector Taylor series for noisy speech recognition. ICSLP'2000, 2000, 869-872

[7] Sankar A, Lee C H. A maximum likelihood approach to stochastic matching for robust speech recognition. *IEEE Trans Speech Audio Processing*, 1996, 4(3): 190-202

[8] Zhu X L. Fundamentals of applied information theory. Beijing: Tsinghua University Press, 2001

[9] Ding C, He X. Cluster merging and splitting in hierarchical clustering algorithms. In: International Conference on Data Mining, 2002: 139-146

[10] Pearce D, Hirsch H. The AURORA experimental framework for the performance evaluation of speech recognition under noisy conditions. In: International Conference on Spoken Language Processing, 2000: 29-32

Zhao Xianyu, born in 1976, is a Ph.D student of Tsinghua University now with major in electronics engineering. His main research interests include intelligent signal processing, robust speech recognition and machine learning.