Stochastic vector mapping-based feature enhancement using prior-models and model adaptation for noisy speech recognition

Chia-Hsin Hsieh, Chung-Hsien Wu *

Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, ROC

Received 29 November 2006; received in revised form 30 January 2008; accepted 4 February 2008

Abstract

This paper presents an approach to feature enhancement for noisy speech recognition. Three prior-models are introduced to characterize clean speech, noise and noisy speech, respectively. Sequential noise estimation is employed for prior-model construction based on noise-normalized stochastic vector mapping. Therefore, feature enhancement can work without stereo training data and manual tagging of background noise type based on the auto-clustering on the estimated noise data. Environment model adaptation is also adopted to reduce the mismatch between training data and test data. For the evaluation on the AURORA2 database, the experimental results indicate that a 9.6% relative reduction in digit error rate for multi-condition training and a 3.5% relative reduction in digit error rate for clean speech training were achieved without stereo training data compared to the SPLICE-based approach. For MATBN Mandarin broadcast news database with multi-condition training, a 13% relative reduction in syllable error rate for anchor speech, a 12% relative reduction in syllable error rate for field reporter speech and a 7% relative reduction in syllable error rate for interviewee speech were obtained compared to the MCE-based approach.

© 2008 Elsevier B.V. All rights reserved.

Keywords: Noisy speech recognition; Cepstral feature enhancement; Environment model adaptation

1. Introduction

State-of-the-art speech recognizers can achieve very high recognition rates under a clean environment, while the recognition rate generally degrades drastically under noisy environment. Noise-robust speech recognition has become an important task for noisy speech recognition. Recent research on noise-robust speech recognition has mostly focused on three directions: (1) noise filtering from the corrupted noisy signal in signal space or feature space. In this direction, spectral subtraction (Boll, 1979), RASTA filtering (Hermansky and Morgan, 1994), Wiener filtering (Macho et al., 2002) and a speech enhancement-based approach (Chen and Wu, 2002) are generally used; (2) speech or feature enhancement using a model-based approach, such as stochastic vector mapping (Wu and Huo, 2002) and SPLICE (Deng et al., 2003) and (3) noise compensation of the acoustic model in the model space so that the training environment can match the test environment. This direction includes PMC (Gales and Young, 1996) and multi-condition/multi-style training (Hirsch and Pearce, 2000; Deng et al., 2000; Yao et al., 2004) methods. The noise filtering approaches require the assumption of prior information, such as the spectral characteristic of the noise. The performance will degrade when the noisy environments vary drastically or under an unknown noise environment. Furthermore, Deng et al. (2003) have shown that denoising or preprocessing is superior to retraining the recognizers under the matched noise conditions with no preprocessing.

Stochastic vector mapping (SVM) (Wu and Huo, 2002; Deng et al., 2003) and sequential noise estimation (Benveniste et al., 1990; Weinstein et al., 1990; Deng et al., 2003) for noise normalization have been proposed and have achieved significant improvements in noisy speech recognition.
However, there still exist some drawbacks and limitations. First, error propagation of the sequential noise estimation generally occurs due to the fixed initial-value setting. The performance of sequential noise estimation will decrease when the noisy environments vary drastically. Second, the environment mismatch between training data and test data still exists and results in performance degradation. Third, the use of maximum-likelihood-based stochastic vector mapping (SPLICE) requires annotation of the environment type and stereo training data. However, the stereo data are not available for most noisy environments.

In this study, in order to minimize error propagation of the fixed initial-value setting in the sequential expectation-maximization (EM) algorithm, the prior-models are introduced to provide more information in the sequential noise estimation. Furthermore, an environment model adaptation using MAP adaptation (Gauvain and Lee, 1994) is constructed to reduce the mismatch between the training data and test data. An MCE-based approach (Wu and Huo, 2002) is employed so that the SVM can work without the stereo training data. Furthermore, an unsupervised frame-based auto-clustering is adopted to automatically detect the environment type of the training data so that the manual classification or tagging is not necessary.

The rest of this study is organized as follows. Section 2 describes the noise-normalized stochastic vector mapping-based cepstral feature enhancement framework. Section 3 then gives the use of prior-model in sequential noise estimation. Section 4 presents experiments on the performance evaluation and discussions of the proposed scheme. Finally, conclusions are given in Section 5.

2. Noise-normalized stochastic vector mapping for cepstral feature enhancement

2.1. Stochastic vector mapping

Fig. 1 shows the framework of the SVM-based feature enhancement approach in the training and testing phases. The SVM-based feature enhancement approach estimates the clean speech feature $\hat{x}$ from noisy speech feature $y$ through an environment-dependent mapping function $F(y; \Theta^e)$ in the cepstral domain, where $\Theta^e$ denotes the correction vectors and $e$ denotes the corresponding environment of noisy speech feature $y$.

Assuming that the training data for the noisy speech $Y$ can be partitioned into $N_E$ different noisy environments, the feature vectors of $Y$ under an environment $e$ can be modeled by a Gaussian mixture model (GMM) with $N_K$ mixtures:

$$p(y|e; \Omega_e) = \sum_{k=1}^{N_E} p(k|e)p(y|k, e) = \sum_{k=1}^{N_K} \alpha_k^e \cdot N(y; \zeta_k^e, \Sigma_k^e),$$

(1)

where $\Omega_e$ represents the environment model. The clean speech feature $\hat{x}$ can be estimated using the stochastic vector mapping function which is defined as follows:

$$\hat{x} \triangleq F(y; \Theta) = y + \sum_{e=1}^{N_E} \sum_{k=1}^{N_K} p(k|y, e)\zeta_k^e,$$

(2)

where the posterior probability $p(k|y, e)$ can be estimated using the Bayes theory based on the environment model $\Omega_e$ as follows:

![Fig. 1. Diagram of the training and testing phase. In the training phase, the noise components are estimated from three pre-trained prior-models and then clustered to train the correction vector for speech enhancement. In the test phase, the input noisy speech is enhanced based on the estimated noise components for speech recognition.](image-url)
and \( \Theta = \{ r^e_k \} \) denotes the correction vectors. Generally, correction vectors \( \Theta \) are estimated from a set of training data using a maximum-likelihood criterion. Fig. 2 illustrates the concept of SVM-based feature enhancement.

For the estimation of the correction vectors \( \Theta \), if stereo data, which contain a clean speech signal and a corrupted noisy speech signal with the identical clean speech signal, are available, the SPLICE-based approach can be directly adopted. However, stereo data are not easily available in real-life applications. This study employs an MCE-based approach to overcome the limitation. In (Wu and Huo, 2002), a standard MCE-based criterion was proposed to estimate the correction vectors and the hidden Markov model (HMM) as follows:

\[
\Theta_{j+1} = \Theta_j - \epsilon_j V_j \nabla l(F(Y_j; \Theta); A)|_{\Theta = \Theta_j};
\]

where \( l(F(Y_j; \Theta); A) \) denotes the loss function under ASR model \( A \).

For the training of the correction vectors \( \Theta \), the environment type of the noisy speech data is needed. The noisy speech data are manually classified into \( N_E \) noisy environment types. This strategy assigns each noisy speech file to only one environment type and is very time consuming. Actually, each noisy speech file contains several segments with different types of noisy environment. Since the noisy speech annotation affects the purity of the training data, with different types of noisy environment. Since the noisy speech is required first. Then the sequential EM algorithm is introduced for online noise estimation based on the relation. In the meantime, the prior-models are invoked to provide more information for noise estimation.

3. Prior-model for sequential noise estimation

The nonlinear acoustic environment model is introduced first for noise estimation in (Deng et al., 2003; Moreno et al., 1996). Given the cepstral features of a clean speech \( x \), an additive noise \( n \) and a channel distortion \( h \), the approximated nonlinear relation among \( x, n, h, \) and the corrupted noisy speech \( y \) in the cepstral domain is estimated as

\[
y \approx h + x + g(n - h - x), \quad g(z) = C \ln(I + \exp[C^T(z)]),
\]

where \( C \) denotes the discrete cosine transform matrix. In order to linearize the nonlinear model, the first-order Taylor series expansion is used around two updated operating points \( n_0 \) and \( \mu_0 \) denoting the initial noise feature and the mean vector of prior clean speech model, respectively. By ignoring the channel distortion effect, for which \( h = 0 \), Eq. (5) is then derived as

\[
y \approx x + g(n_0 - \mu_0) + G(n_0 - \mu_0)(x - \mu_0) + [I - G(n_0 - \mu_0)](n - n_0),
\]

where \( G(z) = -C \text{diag}\left(\frac{I}{1+\exp[C^T(z)]}\right)C^T \).
3.2. The prior-models

The three prior-models $\Phi_n$, $\Phi_x$ and $\Phi_y$, which denote noise, clean speech and noisy speech models, respectively, can provide more information for sequential noise estimation. Fig. 3 illustrates the construction and the use of the three prior-models. First, the noise and clean speech prior-models are characterized by the GMMs as

$$\begin{align*}
\text{prior clean speech model} & \quad \text{prior noise model} \\
\text{prior noisy speech model} & \quad \text{The approximated linear model}
\end{align*}$$

As shown in Fig. 3, the noisy speech prior-model can be employed to search the most similar clean speech mixture component and noise mixture component in sequential noise estimation.

3.3. Sequential noise estimation

In (Deng et al., 2003; Benveniste et al., 1990; Weinstein et al., 1990), a sequential EM algorithm was employed for sequential noise estimation. In this study, the prior clean speech, noise and noisy speech models are used to construct a robust noise estimation procedure.

Based on the sequential EM algorithm, the noise is estimated as $n_{t+1} = \arg \max Q_{t+1}(n)$. In the $E$-step of the sequential EM algorithm, an objective function is defined as

$$Q_{t+1}(n) = E[\ln p(y_{t+1}^n, M_{t+1}^n, D_{t+1}^n | \gamma_{t+1}^n, n_t^1)], \quad (10)$$

where $M_{t+1}^n$ and $D_{t+1}^n$ denote the mixture index sequence of the clean speech GMM and noise GMM in which the noisy speech $y$ occurs from frames 1 to $t+1$. The objective function is simplified for the $M$-step as

$$Q_{t+1}(n) = E[\ln p(y_{t+1}^n, M_{t+1}^n, D_{t+1}^n | \gamma_{t+1}^n, n_t^1) + \ln p(M_{t+1}^n | D_{t+1}^n, n) + \ln p(D_{t+1}^n | n)]$$

where $\gamma_{t+1}^n$ denotes the conditional posterior probability of mixture $n$.
where the likelihood \( p(y_t|m,d,n_{t-1}) \) can be approximated using the approximated linear model (Eq. (6)) as
\[
p(y_t|m,d,n_{t-1}) \sim N[y_t; \mu^x_{m,d}(n_{t-1}), \Sigma_{m,d}^x],
\]
\[
\mu^x_{m,d}(n_{t-1}) = \mu^e_{m,d} + G(n_0 - \mu^e_{m,d})(\mu^e_{m,d} - \mu^o_0) + [I - G(n_0 - \mu^o_0)][n_{t-1} - n_0],
\]
\[
\Sigma_{m,d}^x = [I + G(n_0 - \mu^o_0)]\Sigma_{m,d}[I + G^T(n_0 - \mu^o_0)]^T + [I - G(n_0 - \mu^o_0)]\Sigma_{m,d}[I - G^T(n_0 - \mu^o_0)]^T.
\]
(13)

Also, a forgetting factor is employed to control the effect of the features of the preceding frames:
\[
Q_{t+1}(n) = \sum_{t=1}^{t+1-1} \sum_{m=1}^{N_m} \sum_{d=1}^{N_d} \gamma_t(m,d) \cdot (\ln p(y_t|m,d,n) + \ln p(d_t|n)/(N_m \cdot N_d)) + \text{Const},
\]
\[
\tilde{Q}_{t+1}(n) = - \sum_{t=1}^{t+1-1} \sum_{m=1}^{N_m} \sum_{d=1}^{N_d} \gamma_t(m,d) \cdot ([y_t - \mu^x_{m,d}(n_t)] + \ln p(d_t|n)/(N_m \cdot N_d))
\times \left( \Sigma_{m,d}^x \right)^{-1} [y_t - \mu^x_{m,d}(n_t)] + \ln p(d_t|n)/(N_m \cdot N_d)
\]
\[
R_{t+1}(n) = \sum_{m=1}^{N_m} \sum_{d=1}^{N_d} \gamma_{t+1}(m,d) \cdot ([y_{t+1} - \mu^x_{m,d}(n_{t+1})] + \ln p(d_{t+1}|n)/(N_m \cdot N_d))
\times \left( \Sigma_{m,d}^x \right)^{-1} [y_{t+1} - \mu^x_{m,d}(n_{t+1})] + \ln p(d_{t+1}|n)/(N_m \cdot N_d).
\]
(14)

In the M-step, the iterative stochastic approximation (Benveniste et al., 1990; Krishnamurthy and Moore, 1993) is introduced to derive the solution. Finally, sequential noise estimation is performed as follows:
\[
n_{t+1} = n_t + (K_{t+1})^{-1}s_{t+1}, \quad K_{t+1} = -\frac{\partial^2 Q_{t+1}}{\partial^2 n}
\big|_{n=n_t},
\]
\[
s_{t+1} = -\frac{\partial R_{t+1}}{\partial n}
\big|_{n=n_t}.
\]
(15)

Fig. 4 shows a diagram of the sequential noise estimation. First, the noisy speech prior-model is used to search for the most similar noisy speech mixture component. Then the corresponding noise and clean speech mixture components are determined. Given the two mixture components, the estimation of the posterior probability \( \gamma_t(m,d) \) will be more precise. Furthermore, Fig. 5 shows the relation between the noisy speech, noise, and clean speech features. The first 20 ms of the noisy speech is regarded as the first noise frame.

4. Environment model adaptation

Because the prior-models are usually insufficient for representing the universal data, the environment mismatch between the training data and test data will result in degradation in the performance of the feature enhancement. In this study, an environment model adaptation strategy is proposed for dealing with the problem before the testing phase. Fig. 6 shows a flowchart for the environment model adaptation. The environment model adaptation procedure contains two parts: The first is a model-parameter adaptation to the noise prior-models \( \Phi_y \) and noisy speech prior-model \( \Phi_x \) used in both the training and adaptation phase.
The estimated clean speech components are used to adapt the correction vectors and the environment model.

The second is the adaptation to the noise-normalized SVM function $\Theta$ and environment model $\Omega_e$ (used only in the adaptation phase).

4.1. Model adaptation on noise and noisy speech prior-models

For noise and noisy speech prior-model adaptation, MAP adaptation (Gauvain and Lee, 1994) is first applied to the noise prior-model $\Phi_n$. Given $T$ frames of the adaptation noise data $\{z\}$ estimated from the un-adapted prior-models, the adaptation equation for the noise prior-model parameters is defined as

$$w_d = \frac{(v_d - 1) + \sum_{t=1}^{T} s_{d,t}}{(v_d - 1) + \sum_{d=1}^{N_d} \sum_{t=1}^{T} s_{d,t}},$$

$$\nu_d = \frac{\tau_d p_d + \sum_{t=1}^{T} s_{d,t} \cdot z_t}{\tau_d + \sum_{t=1}^{T} s_{d,t}},$$

$$\Sigma_d^{-1} = \left( v_d + \sum_{t=1}^{T} s_{d,t} (z_t - \bar{\nu}_d) (z_t - \bar{\nu}_d)^T \right) + \tau_d (\bar{\rho}_d - \bar{\nu}_d) (\bar{\rho}_d - \bar{\nu}_d)^T / (\bar{\rho}_d - p) + \sum_{t=1}^{T} s_{d,t},$$

where the conjugate prior density of the mixture weight is the Dirichlet distribution with hyper-parameter $\nu_d$ and the joint conjugate prior density of the mean and variance parameters is the Normal–Wishart distribution with hyper-parameters $\tau_d$, $\rho_d$, $\bar{x}_d$, and $\bar{v}_d$. The two probability distributions are defined as follows (Gauvain and Lee, 1994):

$$g(w_1, \ldots, w_N | v_1, \ldots, v_N) \propto \prod_{d=1}^{N_d} w_d^{v_d - 1},$$

$$g(\mu_d, \Sigma_d | \tau_d, \rho_d, x_d, v_d) \propto |\Sigma_d|^{(p - 1)/2} \exp \left[ -\frac{\tau_d}{2} (\mu_d - \rho_d)^T \tau_d (\mu_d - \rho_d) \right] \times \exp \left[ -\frac{1}{2} \text{tr}(\Sigma_d) \right],$$

where $v_d > 0$, $x_d > p - 1$ and $\tau_d > 0$. After adaptation of the noise prior-model, the noisy speech prior-model $\Phi_s$ is then adapted using the clean speech prior-model $\Phi_s$ and the newly adapted noise prior-model $\Phi_n$ based on Eq. (8).

4.2. Model adaptation for noise-normalized stochastic vector mapping

For noise-normalized SVM adaptation, the environment model parameters $\Omega_e$ and correction vectors in $F(y; \Theta)$ need to be adapted. First, adaptation of the environment model parameter $\Omega_e$ is similar to that of noise prior-model. Second, the adaptation of $\Theta = \{\tau^*\}$ is an iterative procedure. While $\Theta = \{\tau^*\}$ is not a random variable and does not follow any conjugate prior density, an ML-based adaptation which is similar to the correction vector estimation of SPLICE (Deng et al., 2003) is employed:

$$\tilde{\tau}^*_k = \sum_i p(k|y_i - \bar{n}, e) (\bar{x}_i - y_i) \sum_i p(k|y_i - \bar{n}, e),$$

where the temporally estimated clean speech $\bar{x}$ are estimated using the un-adapted noise-normalized stochastic mapping function in Eq. (4).

5. Experimental results and discussion

5.1. Training, adaptation and test sets

In this study, two corpora were introduced for evaluation. The first database is the famous benchmark – the AURORA2 database (Hirsch and Pearce, 2000) for evaluating speaker independent digital speech recognition under a noisy environment. The noise environment of test set $A$ is
similar to that of the training set – including subway, babble, car and exhibition-hall background noises, each at six SNRs: −5 dB to 20 dB in steps of 5 dB, filtered with a G712 channel characteristic. The noisy environment of test set \( B \) is different from that of the training set. Test set \( B \) includes restaurant, street, airport and train station background noises, each at the same range of SNRs and the same channel characteristic as in test set \( A \). Test set \( C \) includes subway and street, each at the same range of SNRs as in test set \( A \), but filtered with an MIRS channel characteristic. The frequency characteristics of telecommunications equipment are simulated for all test conditions. The G712 filtering is used to simulate the frequency characteristics of the GSM terminals, and the MIRS filtering is used to simulate the telecommunications transmission equipment. The training data for clean and multi-condition training comprise 4.96 h with 8440 utterances. Test set \( A \) and test set \( B \) contain 13.81 h with 28,028 utterances, and test set \( C \) comprises 6.90 h with 14,014 utterances. One-fifth of the default test data were used for adaptation. The HTK speech recognizer was utilized and its digit error rate and relative reduction of digit error rate were used to measure the performance.

In addition to AURORA2, the MATBN corpus (Wang et al., 2005) which consists of Mandarin TV News was employed. The news content was collected from 2001 to 2003 and contains 198 h of news shows in total. There are three foreground scene types which were used in the experiment: anchor, field reporter and interviewee. Table 1 shows the statistics of the training, adaptation and test sets. The training set contains 16.38-h speech data which is extracted from the 40-h shows in 2001. The adaptation set comprises 1.15-h speech data which is extracted from five shows: 2003/01/24, 2003/01/27, 2003/02/07, 2003/03/05 and 2003/03/06. The test set consists of 2.04-h speech data which is extracted from another five shows: 2003/01/28, 2003/01/29, 2003/02/11, 2003/03/07 and 2003/04/03. The training data is used to train the prior clean speech model, noisy speech model, noise model, environment model, correction vector and acoustic model. The adaptation data is used for unsupervised adaptation that can adapt the prior noisy speech model, noise model, environment model and correction vector (except acoustic models) using a small set of adaptation data without transcription or tagging, so that the four models can match the test data. The difference between the training and adaptation data is that the recording epoch of adaptation data is nearby the test data and the adaptation data does not have transcriptions. The speech recognizer (Chen et al., 2002; Wu and Yan, 2004) was constructed without any language model, and syllable error rate and relative reduction syllable error rate were used to measure its performance. Furthermore, the average RMS errors between enhanced clean speech and true clean speech feature vectors (the C0 in MFCC feature) on the AURORA2 corpus were also presented. Because this study deals with the feature enhancement in the cepstral domain, the SNR (signal-to-noise ratio) evaluation which is used in the time domain is not presented here.

5.2. Experiments on the AURORA2

Table 2 shows the experimental results of the proposed approach on the AURORA2 database. Two results of previous research are referenced for comparison and three experiments were conducted for different experimental conditions: no denoising (Hirsch and Pearce, 2000), SPLICE with recursive EM using stereo data (Deng et al., 2003), the proposed approach using manual annotation without adaptation, and the proposed approach using auto-clustered training data without and with adaptation. The overall results show that the proposed approach under the lack of stereo training data and manual annotation can achieve a 9.6% relative reduction in digit error rate for multi-condition training and a 3.5% relative reduction in digit error rate for clean speech training compared to the SPLICE-based approach with a recursive EM algorithm and stereo training data.

It is obvious that the proposed approach with auto-clustering achieved a 7.3% and 11.2% relative reduction of digit error rate compared to the proposed approach with manual tagging and SPLICE-based approach. So that the proposed approach with auto-clustering is more self-contained and applicable to practical recognition settings. By comparing to the result between the proposed approach without adaptation and with adaptation, the result with adaptation achieves a 1.2% relative reduction of digit error rate in Set B (in different background noise types to the training data) and a 1.1% relative reduction of digit error rate in Set C (in different background noise types and channel characteristic to the training data), therefore, the environment model adaptation can slightly reduce the mismatch between the training and test data. Table 3 shows the average RMS errors between the enhanced and the true clean speech

<table>
<thead>
<tr>
<th>Scene</th>
<th>Background noise type</th>
<th>Overall</th>
<th>Clean</th>
<th>Speech</th>
<th>Music</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>Anchor</td>
<td>13688.3 s</td>
<td>173.1 s</td>
<td>1702.4 s</td>
<td>414.3 s</td>
<td>4.44 h</td>
</tr>
<tr>
<td></td>
<td>Field reporter</td>
<td>7581.4 s</td>
<td>679.3 s</td>
<td>3343.3 s</td>
<td>17363.5 s</td>
<td>9.74 h</td>
</tr>
<tr>
<td></td>
<td>Interviewee</td>
<td>4324.4 s</td>
<td>268.7 s</td>
<td>636.9 s</td>
<td>2697.0 s</td>
<td>2.20 h</td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>7.11 h</td>
<td>2.01 h</td>
<td>1.58 h</td>
<td>5.69 h</td>
<td>16.38 h</td>
</tr>
<tr>
<td>Adaptation set</td>
<td>Anchor</td>
<td>935.8 s</td>
<td>15.9 s</td>
<td>125.9 s</td>
<td>13.4 s</td>
<td>0.30 h</td>
</tr>
<tr>
<td></td>
<td>Field reporter</td>
<td>37.3 s</td>
<td>738.9 s</td>
<td>360.3 s</td>
<td>1596.6 s</td>
<td>0.76 h</td>
</tr>
<tr>
<td></td>
<td>Interviewee</td>
<td>7.9 s</td>
<td>42.9 s</td>
<td>25.9 s</td>
<td>255.2 s</td>
<td>0.09 h</td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>0.27 h</td>
<td>0.22 h</td>
<td>0.14 h</td>
<td>0.52 h</td>
<td>1.15 h</td>
</tr>
<tr>
<td>Test set</td>
<td>Anchor</td>
<td>1888.6 s</td>
<td>15.6 s</td>
<td>213.5 s</td>
<td>33.6 s</td>
<td>0.59 h</td>
</tr>
<tr>
<td></td>
<td>Field reporter</td>
<td>81.1 s</td>
<td>1330.9 s</td>
<td>642.5 s</td>
<td>2697.3 s</td>
<td>1.32 h</td>
</tr>
<tr>
<td></td>
<td>Interviewee</td>
<td>48.1 s</td>
<td>31.8 s</td>
<td>34.3 s</td>
<td>328.2 s</td>
<td>0.13 h</td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>0.56 h</td>
<td>0.38 h</td>
<td>0.25 h</td>
<td>0.85 h</td>
<td>2.04 h</td>
</tr>
</tbody>
</table>

The scene denotes the foreground speaker type.
feature vector (C0) and the results also show the improvement from using auto-clustering and environment adaptation.

5.3. Experiments on MATBN

Table 4 shows the experimental results of the proposed approach compared to the baseline and MCE-based approaches (Wu and Huo, 2002). Because of the lack of stereo data, SPLICE-based approach was not constructed for comparison. Furthermore, since the MATBN database does not contain detailed noisy environment annotation (such as SNR), auto-clustering is required for environment categorization of the training data. The results demonstrate that the proposed approach outperforms the baseline and the MCE-based approaches for anchor, field reporter and interviewee speech data. Since the MATBN database does not provide true clean speech files, the average RMS errors are not presented here. Obviously, the WER of the field reporter speech (36%) and interviewee speech (62%) are still worse than that of the anchor speech (31%). Some phenomena in the database can explain the result: First, the anchor speech is recorded under a quiet studio environment; in contrast, field reporter speech and interviewee speech contain various rigorous noise environments. Second, the field reporter speech contains some pronunciation variation, mispronunciation and speaker variation. Especially, the interviewee sections are usually accented, spontaneous

---

### Table 2

<table>
<thead>
<tr>
<th>Methods</th>
<th>Training-mode</th>
<th>Set A (%)</th>
<th>Set B (%)</th>
<th>Set C (%)</th>
<th>Overall DER (%)</th>
<th>Overall relative reduction DER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No denoising</td>
<td>Multi-condition</td>
<td>12.18</td>
<td>13.73</td>
<td>16.22</td>
<td>13.61</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Clean only</td>
<td>38.66</td>
<td>44.25</td>
<td>33.86</td>
<td>39.94</td>
<td>–</td>
</tr>
<tr>
<td>MCE</td>
<td>Multi-condition</td>
<td>7.08</td>
<td>10.85</td>
<td>9.91</td>
<td>9.15</td>
<td>32.8</td>
</tr>
<tr>
<td></td>
<td>Clean only</td>
<td>12.18</td>
<td>14.66</td>
<td>16.23</td>
<td>13.98</td>
<td>65.0</td>
</tr>
<tr>
<td>SPLICE with recursive-EM</td>
<td>Multi-condition</td>
<td>8.51</td>
<td>10.84</td>
<td>10.38</td>
<td>9.82</td>
<td>27.9</td>
</tr>
<tr>
<td></td>
<td>Clean only</td>
<td>12.18</td>
<td>12.91</td>
<td>14.92</td>
<td>13.02</td>
<td>65.2</td>
</tr>
<tr>
<td>Proposed approach (manual tag, no adaptation)</td>
<td>Multi-condition</td>
<td>8.39</td>
<td>10.61</td>
<td>9.96</td>
<td>9.66</td>
<td>29.0</td>
</tr>
<tr>
<td></td>
<td>Clean only</td>
<td>12.01</td>
<td>12.87</td>
<td>14.60</td>
<td>12.99</td>
<td>67.5</td>
</tr>
<tr>
<td>Proposed approach (auto-clustering, no adaptation)</td>
<td>Multi-condition</td>
<td>8.76</td>
<td>9.03</td>
<td>9.04</td>
<td>8.95</td>
<td>34.2</td>
</tr>
<tr>
<td></td>
<td>Clean only</td>
<td>12.28</td>
<td>12.54</td>
<td>13.51</td>
<td>12.62</td>
<td>68.4</td>
</tr>
<tr>
<td>Proposed approach (auto-clustering, with adaptation)</td>
<td>Multi-condition</td>
<td>8.74</td>
<td>8.92</td>
<td>8.94</td>
<td>8.88</td>
<td>34.8</td>
</tr>
<tr>
<td></td>
<td>Clean only</td>
<td>12.27</td>
<td>12.42</td>
<td>13.45</td>
<td>12.57</td>
<td>68.5</td>
</tr>
</tbody>
</table>

The clean only and multi-condition training denote the acoustic model is trained only on clean speech data and on all speech data in different background noises.

### Table 3

<table>
<thead>
<tr>
<th>Methods</th>
<th>Set A (%)</th>
<th>Set B (%)</th>
<th>Set C (%)</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed approach (manual tag, no adaptation)</td>
<td>3.92</td>
<td>4.51</td>
<td>4.46</td>
<td>4.28</td>
</tr>
<tr>
<td>Proposed approach (auto-clustering, no adaptation)</td>
<td>3.99</td>
<td>4.22</td>
<td>4.21</td>
<td>4.01</td>
</tr>
<tr>
<td>Proposed approach (auto-clustering, with adaptation)</td>
<td>3.96</td>
<td>4.05</td>
<td>4.03</td>
<td>3.99</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods</th>
<th>Scene</th>
<th>Background noise type (%)</th>
<th>Overall SER (%)</th>
<th>Overall relative reduction SER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No denoising</td>
<td>Anchor</td>
<td>Clean 33</td>
<td>Speech 24</td>
<td>Music 58</td>
</tr>
<tr>
<td></td>
<td>Field reporter</td>
<td>43</td>
<td>47</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>Interviewee</td>
<td>63</td>
<td>71</td>
<td>69</td>
</tr>
<tr>
<td>Stochastic vector mapping + MCE</td>
<td>Anchor</td>
<td>Clean 31</td>
<td>Speech 23</td>
<td>Music 56</td>
</tr>
<tr>
<td></td>
<td>Field reporter</td>
<td>38</td>
<td>43</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>Interviewee</td>
<td>59</td>
<td>67</td>
<td>64</td>
</tr>
<tr>
<td>Proposed approach (no adaptation)</td>
<td>Anchor</td>
<td>Clean 29</td>
<td>Speech 21</td>
<td>Music 53</td>
</tr>
<tr>
<td></td>
<td>Field reporter</td>
<td>36</td>
<td>39</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Interviewee</td>
<td>57</td>
<td>65</td>
<td>60</td>
</tr>
<tr>
<td>Proposed approach (with adaptation)</td>
<td>Anchor</td>
<td>Clean 29</td>
<td>Speech 20</td>
<td>Music 52</td>
</tr>
<tr>
<td></td>
<td>Field reporter</td>
<td>35</td>
<td>38</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Interviewee</td>
<td>56</td>
<td>64</td>
<td>59</td>
</tr>
</tbody>
</table>
and disfluent. The cepstral feature enhancement is still not robust enough to overcome the problem.

Furthermore, a special experimental result should be discussed about anchor speech with background music. Since the SNR of anchor speech under background music which usually happens in preview sections is lower than that of anchor speech under background speech and others, the recognition accuracy of the anchor speech under background music is worse than that of anchor speech under other three background noises. Moreover, the recognition accuracy of field reporter speech is better than that of anchor speech under background music. Similar to a human’s behavior, under the low SNR situation, further knowledge (such as a language model) is necessary for large vocabulary speech recognition.

6. Conclusions

This study has presented an approach to cepstral feature enhancement for noisy speech recognition using noise-normalized stochastic vector mapping. The prior-model was introduced for precise noise estimation. Then the environment model adaptation is constructed to reduce the environment mismatch between training data and test data. The experimental results demonstrate that the proposed approach with auto-clustering on the AURORA2 database can slightly outperform the SPLICE-based approach with the stereo data as well as the proposed approach with manual tagging of background noise type. For the Mandarin news corpus MATBN, the proposed approach also achieves satisfactory improvements compared to the baseline and the MCE-based approaches.

References


