Stochastic Vector Mapping-based Feature Enhancement Using Prior Model and Environment Adaptation for Noisy Speech Recognition

Chia-Hsin Hsieh, Chung-Hsien Wu and Jun-Yu Lin

Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan, R.O.C.
{ ngsnail, chwu, adversary}@csie.ncku.edu.tw

10.21437/Interspeech.2006-8

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 using sequential noise estimation based on noise-normalized stochastic vector mapping. Environment adaptation is also adopted to reduce the mismatch between training data and test data. For AURORA2 database, the experimental results indicate that a 0.77% digit accuracy improvement for multi-condition training and 0.29% digit accuracy improvement for clean speech training were achieved without stereo training data compared to the SPLICE-based approach with recursive noise estimation. For MAT-BN Mandarin broadcast news database, a 2.6% syllable accuracy improvement for anchor speech and 4.2% syllable accuracy improvement for field report speech were obtained compared to the MCE-based approach.

Index Terms: noisy speech recognition, feature enhancement, environment adaptation, prior model

1. INTRODUCTION

The state-of-the-art speech recognizers can achieve very high recognition rate under clean environment, while the recognition rate generally degrades drastically under noisy environment. Therefore, noise-robust speech recognition has become an important task for noisy speech recognition. Recent research on noise-robust speech recognition mostly focused on two directions: (1) Remove the noise from the corrupted noisy signal in signal or feature space, such as spectral subtraction, and model-based feature enhancement: SPLICE [1]; (2) Model compensation in model space, such as PMC [7].

The stochastic vector mapping (S.V.M.) [1-2] with sequential noise estimation-based noise normalization [1,3,5] have been proposed and achieved high improvement in noisy speech recognition. However, there still exist some drawbacks and limitations. First, the performance of sequential noise estimation will decrease when the noisy environment vary drastically. Second, the environment mismatch between training data and test data still exists and results in performance degradation. Third, the maximum-likelihood-based stochastic vector mapping (SPLICE) required annotation of environment type and stereo training data. Nevertheless, the stereo data are not available for most noisy environments. In order to overcome the insufficiency of tracking ability in sequential EM, the prior models are introduced to provide more information in sequential noise estimation. Furthermore, an environment adaptation is constructed to reduce the mismatch between the training data and test data. Finally, MCE-based approach [2] is employed without the stereo training data and an unsupervised frame-based auto-clustering is adopted to automatically detect the environment type of the training data.

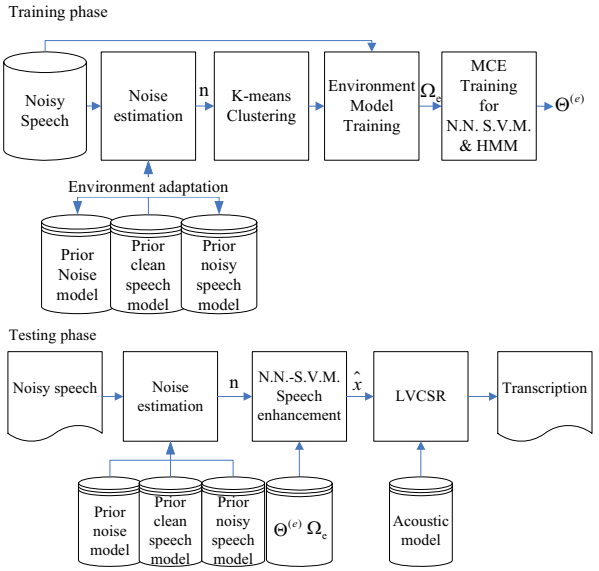

Figure 1: Detailed flowchart of the training and testing phase

2. NOISE-NORMALIZED STOCHASTIC VECTOR MAPPING FOR FEATURE ENHANCEMENT

2.1 Stochastic Vector Mapping (S.V.M.)

Figure 1 shows the frameworks of the proposed S.V.M.-based feature enhancement approach in training, adaptation and testing phases. The S.V.M.-based feature enhancement approach estimates the clean speech feature $\hat{x}$ from noisy speech feature y through an environment-dependent mapping function $F(y;\Theta^{(e)})$, where $\Theta^{(e)}$ denotes the mapping function parameters and e denotes the corresponding environment of the noisy speech y.

Assuming that the training data of the noisy speech $\boldsymbol{Y}$ can be partitioned into K different noisy environments, the feature of $\boldsymbol{Y}$ under an environment e can be modeled by a Gaussian mixture model (GMM):

$$p(y \mid e;\Omega_e) = \sum_{k=1}^{K} p(k \mid e)\, p(y \mid k,e) = \sum_{k=1}^{K} \omega_k^e \cdot N\left(y;\xi_k^e, R_k^e\right) \quad (1)$$

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

$$\hat{x} \triangleq F\left(y;\Theta^{(e)}\right) = y + \sum_{k=1}^{K} p(k \mid y,e)\, r_k^e \quad (2)$$

where $\Theta^{(e)} = \left\{r_k^{(e)}\right\}_{k=1}^{K}$ denotes the mapping function parameters and the posterior probability $p(k \mid y,e)$ can be estimated using the Bayes theory based on the environment model Ω_e as follows:

$$p(k \mid y,e) = p(k \mid e)\, p(y \mid k,e) \Big/ \sum_{j=1}^{K} p(j \mid e)\, p(y \mid j,e) \quad (3)$$

Generally, $\Theta^{(e)}$ are estimated from a set of training data using maximum likelihood criterion. For the estimation of the mapping function parameter $\Theta^{(e)}$, if the stereo data (a clean speech signal and the corrupted noisy speech with the identical clean speech signal) are available, the SPLICE approach can be directly adopted. However, the stereo data are not available in real-life applications. This study employs an MCE-based approach to overcome the limitation. In [2], the MCE-based criterion was proposed to estimate the parameters of the mapping function and the hidden Markov model (HMM). This approach can obtain satisfactory results without the stereo data. Furthermore, the environment type of the noisy speech data is needed for training the environment model Ω_e. Annotation of the environment type for the noisy speech is to roughly classify the noisy speech data into *E* noisy environments manually by listening to the background noises in the speech file. This strategy assigns each noisy speech file to only one environment type and is time consuming. Actually, each noisy speech contains several segments with different types of noisy environment. Since annotation of noisy speech affects the purity of the environment model Ω_e, this study introduces a frame-based unsupervised noise clustering approach to construct a more precise categorization of the noisy speech.

2.2 Noise-Normalized Stochastic Vector Mapping

In [1], the concept of noise normalization is proposed to reduce the effect of background noise in the noisy speech for feature enhancement. If the noise feature vector $\tilde{n}$ of each frame can be estimated first, the noise-normalized stochastic vector mapping (N.N.-S.V.M.) is conducted by replacing y and $\hat{x}$ with $y - \tilde{n}$ and $\hat{x} - \tilde{n}$ as

$$\hat{x} - \tilde{n} \triangleq F\left(y - \tilde{n};\Theta^{(e)}\right) = y - \tilde{n} + \sum_{k=1}^{K} p\left(k \mid y - \tilde{n}, e\right) r_k^e \quad (4)$$

Obviously, the estimation algorithm of noise feature vector $\tilde{n}$ plays an important role in noise-normalized stochastic vector mapping.

3. PRIOR MODEL FOR SEQUENTIAL NOISE ESTIMATION

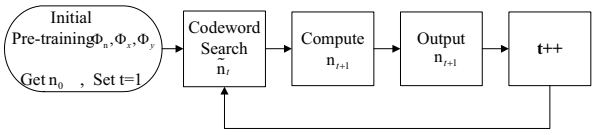

Figure 2: Flowchart of noise estimation

This study employs a frame-based sequential noise estimation algorithm [1,3,5] by incorporating the prior models. Figure 2 shows the flowchart of noise estimation. In the procedure, only noisy speech feature vector is observed in the current frame. Since the noise and clean speech feature vector are missing simultaneously, the relation among clean speech, noise and noisy speech is required first. Then the prior models are constructed to provide more information for noise estimation.

3.1 Acoustic Environment Model

The nonlinear acoustic environment model is introduced first for noise estimation in [1]. Given the cepstral feature of a clean speech x, additive noise n and channel distortion h, the approximated nonlinear relation among x, n, h and the corrupted noisy speech y in cepstral domain is estimated as:

$$y \approx h + x + g(n - h - x), \quad g(z) = C \ln\left(I + \exp\left[C^T(z)\right]\right) \quad (5)$$

where **C** denotes the discrete cosine transform matrix. In order to linearlize the nonlinear model, the first order Taylor series expansion was used around two updated operating points n_0 and μ_0^x. By ignoring the channel distortion effect, for which h=0, Eq. (5) is then derived as:

$$y \approx \mu_0^x + g\left(n_0 - \mu_0^x\right) + G\left(n_0 - \mu_0^x\right)\left(x - \mu_0^x\right) + \left[I - G\left(n_0 - \mu_0^x\right)\right]\left(n - n_0\right) \quad (6)$$

where $G(z) = -C\,\mathrm{diag}\left(I / \left(I + \exp\left[C^T z\right]\right)\right) C^T$.

3.2 Prior Models

The three prior models Φ_x, Φ_n and Φ_y which denote clean speech, noise and noisy speech respectively are required for sequential noise estimation. First, the noise and clean speech prior models are defined as follows:

$$p(n;\Phi_n) = \sum_{c=1}^{C} w_c \cdot N\left(n;\mu_c^n,\Sigma_c^n\right),\ p(x;\Phi_x) = \sum_{m=1}^{M} w_m \cdot N\left(x;\mu_m^x,\Sigma_m^x\right) \quad (7)$$

where pre-training data are required to train the model parameters of the two GMMs Φ_x and Φ_n.

While the prior noisy speech model is needed in sequential noise estimation, the noisy speech model parameters are derived according to the prior clean speech and noise models (the approximated linear model) using Eq. (6) as follows:

$$\begin{aligned} &p(y;\Phi_y) = \sum_{m=1}^{M}\sum_{c=1}^{C} w_{m,c} \cdot N\left(y;\mu_{m,c}^y,\Sigma_{m,c}^y\right) \\ &\mu_{m,c}^y = \mu_0^x + g\left(\mu_0^n - \mu_0^x\right) + G\left(\mu_0^n - \mu_0^x\right)\left(\mu_m^x - \mu_0^x\right) + \left[I - G\left(\mu_0^n - \mu_0^x\right)\right]\left(\mu_c^n - \mu_0^n\right) \\ &\Sigma_{m,c}^y = \left[I + G\left(\mu_0^n - \mu_0^x\right)\right]\Sigma_m^x\left[I + G^T\left(\mu_0^n - \mu_0^x\right)\right]^T \\ &\mu_0^n = E[\mu_c^n],\ \mu_0^x = E[\mu_m^x],\ w_{m,c} = w_m \bullet w_c \end{aligned} \quad (8)$$

3.3 Sequential Noise Estimation

In [1,3,5], sequential expectation-maximization (EM) algorithm is employed for sequential noise estimation. In this study, the prior clean speech, noise and noisy speech model is involved to construct a robust noise estimation procedure.

Based on the sequential EM algorithm, the estimated noise is obtained using $\mathrm{n}_{t+1} = \arg\max_{\mathrm{n}} Q_{t+1}(\mathrm{n})$. In the E-step, an objective function is defined first as:

$$Q_{t+1}(\mathrm{n}) \triangleq E\left[\ln p(y_1^{t+1}, M_1^{t+1}, C_1^{t+1} \mid \mathrm{n}) \mid y_1^{t+1}, \mathrm{n}_1^t\right] \quad (9)$$

where M_1^{t+1} and C_1^{t+1} denotes the mixture indexes of the clean speech model and noise model to which the noisy speech y occurs from frame 1 to t+1. In the M-step, the iterative stochastic approximation and a forgetting factor are introduced. Finally, a sequential noise estimation function is derived.

4. ENVIRONMENT ADAPTATION

Because the prior models are usually not complete enough to represent the universal data, the environment mismatch between training data and test data will result in the degradation on feature enhancement performance. In this study, an environment adaptation strategy is proposed before testing phase to deal with the problem. Figure 3 shows the environment adaptation flowchart. The environment adaptation procedure contains two parts: The first one is model parameter adaptation on noise prior model Φ_n and noisy speech prior model Φ_y and the second is on noise-normalized S.V.M. function $\Theta^{(e)}$ and environment model Ω_e.

4.1 Model Adaptation on Noise and Noisy Speech Prior Models

For noise and noisy speech prior model adaptation, MAP adaptation [8] is applied to the noise prior model Φ_n first. The adaptation equation for the noise prior model parameters given T frames of the adaptation data Z is defined as:

$$\begin{aligned}
\tilde{w}_c &= (v_c - 1) + \sum_{t=1}^{T} d_{c,t} \Big/ \sum_{c=1}^{C} (v_c - 1) + \sum_{c=1}^{C}\sum_{t=1}^{T} d_{c,t} \\
\tilde{\mu}_c^n &= \tau_c \rho_c + \sum_{t=1}^{T} d_{c,t} \cdot y_t \Big/ \tau_c + \sum_{t=1}^{T} d_{c,t} \\
\tilde{\Sigma}_c^{n\,-1} &= \upsilon_c + \sum_{t=1}^{T} d_{c,t} \left(z_t - \tilde{\mu}_c^n\right)\left(z_t - \tilde{\mu}_c^n\right)^T + \tau_c \left(\rho_c - \tilde{\mu}_c^n\right)\left(\rho_c - \tilde{\mu}_c^n\right)^T \Big/ (\alpha_c - p) + \sum_{t=1}^{T} d_{c,t}
\end{aligned} \quad (10)$$

where the conjugate prior density of the mixture weight is the Dirichlet distribution with hyper-parameter of v_c and the joint conjugate prior density of mean and variance parameters is the Normal-Wishart distribution with hyper-parameter of τ_c, ρ_c, α_c, and υ_c.

After adaptation of noise prior model, the noisy speech prior model Φ_y is then adapted using the newly adapted noise prior model Φ_n according to Eq. (8).

4.2 Model Adaptation of Noise Normalized Stochastic Vector Mapping

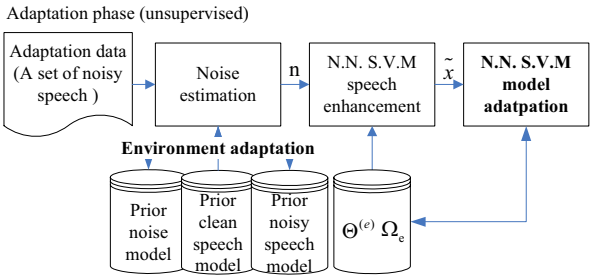

Figure 3: Detailed flowchart of the adaptation phase

For noise-normalized S.V.M. adaptation, model parameters Ω_e and mapping function parameters in $F\left(y;\Theta^{(e)}\right)$ need to be adapted in the adaptation phase and testing phase, respectively. First, adaptation of model parameter Ω_e is similar to that of noise prior model. Second, the adaptation of $\Theta^{(e)} = \left\{r_k^{(e)}\right\}_{k=1}^{K}$ is an iterative procedure. While $\Theta^{(e)} = \left\{r_k^{(e)}\right\}_{k=1}^{K}$ 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 is employed as:

$$\widetilde{r_k^{(e)}} = \sum_t p\left(k \mid y_t - \tilde{n}, e\right)\left(\tilde{x}_t - y_t\right) \Big/ \sum_t p\left(k \mid y_t - \tilde{n}, e\right) \quad (11)$$

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

5. EXPERIMENTAL RESULTS

5.1 Training, Adaptation and Test Sets

In this study, two corpora were introduced for evaluation. The first database is the famous benchmark—AURORA2 database [8] for noisy speech recognition evaluation. One fifth of the default test data were extracted for adaptation. The HTK speech recognizer was introduced and the digit recognition accuracy was used to measure the performance.

The second is the MAT-BN corpus [9] which consists of Mandarin TV News. The news content is collected from anchors, field reporters and interviewers. One hundred and twenty hours news audio were extracted for training, adaptation and testing. The speech recognizer [7] was constructed without any language model and the syllable accuracy was used to measure the performance.

5.2 Experiments on AURORA2

Table 1 shows the experimental results of the proposed approach on AURORA2 database. Two results of previous research were referenced for comparison and three experiments were conducted for different experimental conditions: no denoising (BASELINE) [8], SPLICE with recursive EM using stereo data (SPLICE+R_EM) [1], proposed approach using manual annotation without adaptation (N.N._S.V.M.+MA-AD),

the proposed approach using auto-clustered training data without adaptation (N.N._S.V.M.+ AC–AD) and with adaptation (N.N._S.V.M. +AC+AD). The overall results show that the proposed approach can slightly outperform the SPLICE approach with recursive EM algorithm under the lack of stereo training data and manual annotation. Furthermore, the environment adaptation can reduce the mismatch between training data and test data.

Table 1: Experimental result on AURORA2

Methods	Training Mode	Set A	Set B	Set C	Overall
BASELINE	Multi-cond.	87.82%	86.27%	83.78%	86.39%
	Clean only	61.34%	55.75%	66.14%	60.06%
SPLICE +R_EM	Multi-cond.	91.49%	89.16%	89.62%	90.18%
	Clean only	87.82%	87.09%	85.08%	86.98%
N.N._S.V.M. +MA-AD	Multi-cond	91.42%	89.18%	89.85%	90.21%
	Clean only	87.84%	86.77%	85.23%	86.89%
N.N._S.V.M. +AC–AD	Multi-cond.	91.06%	90.79%	90.77%	90.89%
	Clean only	87.56%	87.33%	86.32%	87.22%
N.N._S.V.M. +AC+AD	Multi-cond.	91.07%	90.90%	90.81%	90.95%
	Clean only	87.55%	87.44%	86.38%	87.27%

5.3 Experiments on MAT-BN

Table 2 shows the experimental results of the proposed approach compared to the baseline and MCE-based approaches [2]. Because of the lack of stereo data, SPLICE-based approach was not constructed for comparison. Furthermore, since MAT-BN 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 the proposed approach outperformed the baseline and MCE-based approaches for both anchor and field report speech data. However, the recognition accuracy of field report speech is still worse than that of anchor speech. This is because the field report speech contains spontaneous and disfluent speech. Cepstral feature enhancement is still not robust enough to overcome the problem.

Table 2: Experimental results on MAT-BN

Speaker	Back-ground	BASE-LINE	MCE	N.N._ S.V.M. –AD	N.N._ S.V.M. +AD
Anchor	Speech	58.1%	59.1%	61.1%	61.9%
	Music	55.7%	57.5%	59.6%	60.8%
	Other	63.9%	68.4%	69.4%	70.1%
	Overall	59.2%	61.7%	63.4%	64.3%
Field report	Speech	27.9%	32.3%	35.8%	37.3%
	Music	21.9%	28.1%	31.2%	32.8%
	Other	30.9%	36.1%	38.0%	39.1%
	Overall	26.9%	32.2%	35.0%	36.4%

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 adaptation is constructed to reduce the environment mismatch between training data and test data. The experimental results demonstrate that the proposed approach can slightly outperform the SPLICE-based approach without stereo data on AURORA2 database. Furthermore, on the Mandarin news corpus, the proposed approach also achieves satisfactory improvements compared to the baseline and MCE-based approaches.

7. REFERENCE

[1] L. Deng, J. Droppo, and Alex Acero, “Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition,” *IEEE Trans. on Speech and Audio Proc.*, vol. 11, No. 6, pp. 568-580, 2003.

[2] J. Wu and Q. Huo, “An environment compensated minimum classification error training approach and its evaluation on AURORA2 database,” in *Proc. ICSLP-2002*, Denver, Colorado, USA, 2002, pp.453-456

[3] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations-Applications of Mathematics*. New York: Springer, 1990. vol.22.

[4] E. Weinstein, M. Feder, and A. Oppenheim, “Sequential algorithms for parameter estimation based on Lullback-Leibler information measure,” *IEEE Trans. on Signal Proc.*, vol. 38, pp. 1652-1654, 1990.

[5] M.J.F. Gales and S.J. Young, “Robust Continuous Speech Recognition Using Parallel Model Combination,” *IEEE Trans. on Speech and Audio Proc.*, vol. 4, no. 5, pp. 352-359, 1996.

[6] Jean-Luc Gauvain and Chin-Hui Lee, “Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chins,” *IEEE Transactions on Speech and Audio Pro.,* Vol. 2, No. 2, 1994.

[7] Chung-Hsien Wu and Gwo-Lang Yan, “Acoustic Feature Analysis and Discriminative Modeling of Filled Pauses for Spontaneous Speech Recognition,” *Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology*, 36, 2004, pp.87-99.

[8] H. Hirsch and D. Pearce, “The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions,” *in Proc. ISCA ITRWASR2000 on Automatic Speech Recognition: Challenges for the Next Millennium*, Paris, France, Sept. 2000.

[9] Hsin-Min Wang, “MATBN 2002: A Mandarin Chinese Broadcast News Corpus” *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition* (SSPR2003)