An Additive and Convolutive Bias Compensation Algorithm for Telephone Speech Recognition[1)]

HAN Zhao-Bing ZHANG Shu-Wu XU Bo HUANG Tai-Yi
(*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100080)
(E-mail: {zbhan, swzhang, xubo, huang}@hitic. ia. ac. cn)

Abstract A Vector piecewise polynomial (VPP) approximation algorithm is proposed for environment compensation of speech signals degraded by both additive and convolutive noises. By investigating the model of the telephone environment, we propose a piecewise polynomial, namely two linear polynomials and a quadratic polynomial, to approximate the environment function precisely. The VPP is applied either to the stationary noise, or to the non-stationary noise. In the first case, the batch EM is used in log-spectral domain; in the second case the recursive EM with iterative stochastic approximation is developed in cepstral domain. Both approaches are based on the minimum mean squared error (MMSE) sense. Experimental results are presented on the application of this approach in improving the performance of Mandarin large vocabulary continuous speech recognition (LVCSR) due to the background noises and different transmission channels (such as fixed telephone line and GSM). The method can reduce the average character error rate (CER) by about 18%.

Key words Speech recognition, piecewise polynomial approximation, environment compensation, recursive EM algorithm

1 Introduction

Basically, the distortion sources in telephone network fall into two categories: 1) noise contamination including background noise and electrical noise, and 2) channel effect caused by transmission line and telephone handset. Due to these distortion sources, the telephone speech recognition performance is seriously damaged[1].

In the recent years, many methods have been proposed to compensate the noisy effect. Accordingly, the Codeword Dependent Cepstral Normalization method (CDCN)[2], the Multivariate Gaussian Based Cepstral Normalization (RATZ) and a Vector Taylor Series (VTS)[3] were presented for reducing the variability of additive noise and channel effect. These approaches use the MMSE estimator to model the noisy speech as a clean signal plus a correction vector. In CDCN, the correction vector is a weighted sum of codeword-dependent correction. However, the noise only depends on the first codebook and the correction vector is constant for all vectors in a codebook. Although the correction of RATZ is a weighted sum of multivariate Gaussian mixture and it does not have the above two disadvantages, the correction does not represent the underlying structure of the noise environment. VTS uses an environment function to describe the correction vector in detail and analytically approximates the function by the truncated Taylor series expansion around the mean of the clean speech. But, it is known that Taylor series is not a precise approximation if the distribution of the variable is not around the point of Taylor′s expansion.

In this paper, we propose a vector piecewise polynomial method, in which we approximate the environment function piecewisely (two linear polynomials and a quadratic polynomial). By taking account of the property of the environment function, the approxima-

[1)] Supported by National Natural Science Foundation of P. R. China (69835003) and the National "973" Plan (G1998030504)
Received December 2, 2002; in revised form July 25, 2003
收稿日期 2002-12-02; 收修改稿件日期 2003-07-25

tion using VPP method is very close to the actual function. Furthermore, the approach does not so strongly rely on the expansion point as VTS. In addition, an HMM with Gaussian output is exploited to classify the acoustic space of clean speech.

The paper is organized as follows. First the property of the environment function is analyzed. We then explain how to use the piecewise polynomial to approximate it. After that, the environment estimation framework of the stationary and non-stationary noise using VPP is introduced. And finally, the experimental results are given and some experimental conclusions are drawn.

2 VPP approximation algorithm

2.1 Model of the environment

As we all know, a commonly accepted model in the log-spectral domain [2] is:

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{h} + \log(1 + \mathrm{e}^{\boldsymbol{n}-\boldsymbol{x}-\boldsymbol{h}}) \tag{1}$$

Or in more general terms:

$$\boldsymbol{y} = \boldsymbol{x} + f(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{h})$$

where $\boldsymbol{h}$ is an unknown parameter that represents the effect of linear filter and $\boldsymbol{n}$ represents the effect of background noise. $f(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{h})$ denotes the environment function.

To simplify the notation, we define the vector function $g(\boldsymbol{v})$ as

$$g(\boldsymbol{v}) = f(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{h}) - \boldsymbol{h} = \log(1 + \exp(\boldsymbol{v})) \tag{2}$$

$g(\boldsymbol{v})$ is a nonlinear function. VTS uses the Taylor series expansion around the operating points ($\boldsymbol{\mu}_x$, $\boldsymbol{n}_0$ and $\boldsymbol{h}_0$) to approximate $g(\boldsymbol{v})$. The accuracy of this approximation strongly depends on the initial values or the operating points.

2.2 The property of $g(v)$

From Fig. 1, $g(v)$ is monotonically increasing, with asymptotes at $g(v) = 0$ for $v \to -\infty$ and $g(v) = v$ for $v \to +\infty$. In the ranges $(-\infty, a)$ and $(b, +\infty)$ the plot of $g(v)$ is very close to a straight line; in the range (a, b) , the plot is a curve.

In (3), $h(\boldsymbol{v})$ represents the derivative of the vector function.

$$h(\boldsymbol{v}) = \frac{\mathrm{d}g(\boldsymbol{v})}{\mathrm{d}\boldsymbol{v}} = \frac{\mathrm{e}^{\boldsymbol{v}}}{1 + \mathrm{e}^{\boldsymbol{v}}} \tag{3}$$

From Fig. 2, in the range (a, b), $h(v)$ is nearly a straight line. Therefore, we adopt a quadratic polynomial to approximate $g(v)$. The constants (a, b) can be empirically determined with an acceptable precision. Additionally in our experiment it proves that a quadratic polynomial is enough to approximate $g(v)$ in the range (a, b), instead of a cubic or higher order polynomial.

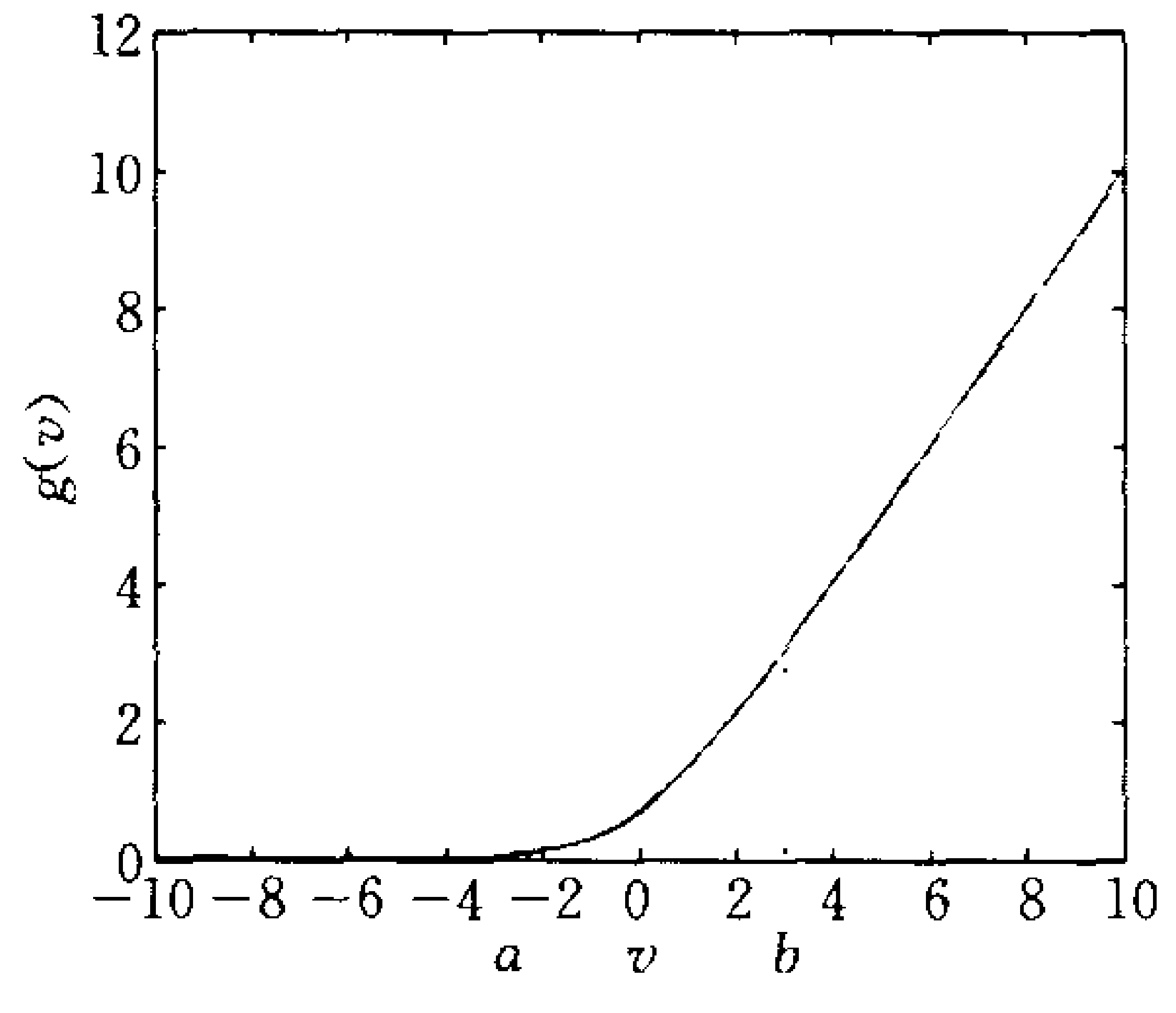

Fig. 1 The function $g(v) = \log(1 + \exp(v))$

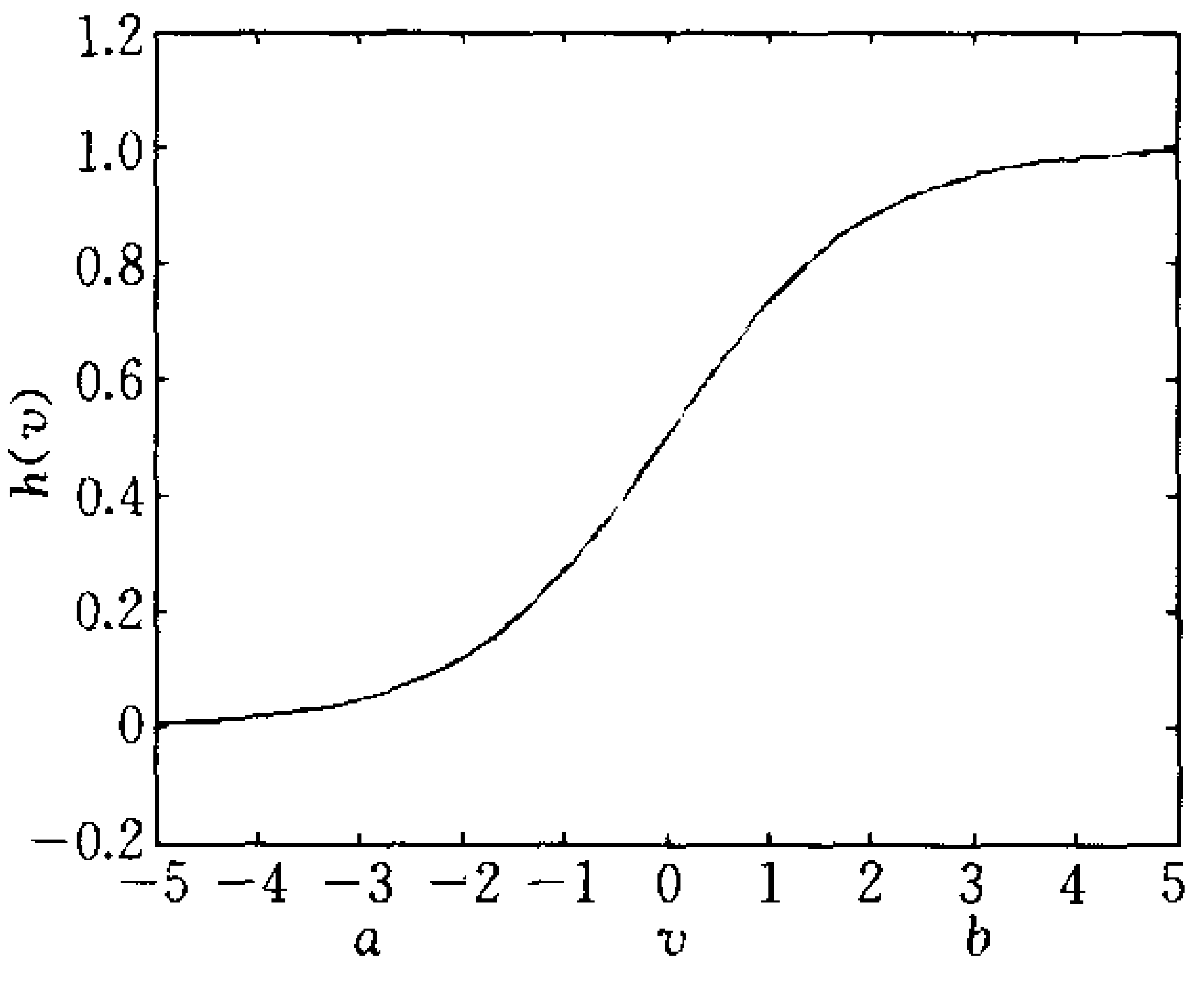

Fig. 2 The function $h(v) = \exp(v)/(1 + \exp(v))$

2.3 The piecewise polynomial approximation

In our approach, linear approximation of $g(v)$ in the range of $(-\infty, a)$ and $(b, +\infty)$

is obtained by minimizing the mean squared error. The range of (a,b) is replaced by a quadratic Chebyshev Polynomial.

The function prototype of the linear function is

$$g(v) = Av + B$$

And the quadratic function is

$$g(v) = Av^2 + Bv + C$$

where A, B and C are constants and determined beforehand.

In the previous work[1], it was shown that it is reasonable to adopt the Gaussian assumption for clean and noisy speech. This assumption implies a linear transformation between the log spectra of clean and noisy speech for each Gaussian density component of the PDF of clean speech. Therefore, to estimate the noise and channel parameters, we approximate the environment function (Eq. (1)) by the statistical linear transformation:

$$\boldsymbol{y} \cong (1-A)\boldsymbol{x} + (1-A)\boldsymbol{h} + A\boldsymbol{n} + B \tag{4}$$

So the estimates of mean and variance of the Kth Gaussian of the noisy speech can be gotten as functions of A, B:

$$\mu_{y,k} = (1-A_k)(\mu_{x,k} + h) + A_k\mu_n + B_k \tag{5}$$

$$\sigma^2_{y,k} = (1-A_k)^2\sigma^2_{x,k} + A^2_k\sigma^2_n \tag{6}$$

Since the estimation of $\mu_{y,k}$ and $\sigma^2_{y,k}$ can be obtained from the VPP algorithm. From (4) and (5), we can compute A_k and B_k. Therefore,μ_n, σ_n and h will be derived from EM algorithm. After that, we use the MMSE criterion to calculate the clean speech, given the observed noisy speech.

Compared with VTS, VPP makes use of the HMM parameters and does not strictly depend on the initial values. The Vector Polynomial approximation S (VPS) algorithm (B. Raj, 1996)[5] is analogous to VPP. But VPS uses the triangular function to approximate the second derivative of environment function. Although it may be very exact to approximate the derivative, its approximation of the function $g(v)$ which is derived from integrating the derivative twice may be not precise. Moreover, it adds the computational complexity.

3 The environment estimation framework

In this section, we first apply the VPP algorithm to estimate the stationary noise and telephone channel with the batch EM. Then the VPP based estimation of the non-stationary noise and channel using the recursive EM is proposed. The formulas of both estimation methods are derived, respectively.

3.1 The estimation of stationary noise and channel

In this case, the optimal parameter is $\lambda = \{\mu_n, \Sigma_n, \boldsymbol{h}\}$. The auxiliary function, Q is defined by

$$Q(\lambda,\bar{\lambda}) = E[\log p(\boldsymbol{X},\boldsymbol{N},K \mid \bar{\lambda} = \{\mu_n,\Sigma_n,\boldsymbol{h}\}) \mid \boldsymbol{y},\lambda = \{\mu_n,\Sigma_n,\boldsymbol{h}\}]$$

where $\boldsymbol{N} = \{\boldsymbol{n}_1, \boldsymbol{n}_2, \cdots, \boldsymbol{n}_T\}$ denotes the noise vector sequence which is statistically independent of the clean feature vector sequence $\boldsymbol{X}$ in log-spectral domain, and $K = \{k_1, k_2, \cdots, k_T\}$ is a hidden sequence of mixture components. By following the way similar to that used in [6], we can obtain the formulas of the estimation of the environment parameters:

$$\hat{\mu}_n = \frac{\sum_{t=1}^{T}\sum_{k=1}^{K} P(k \mid \boldsymbol{y}_t, \lambda)\, \mu_n(\boldsymbol{y}_t, k, \lambda)}{\sum_{t=1}^{T}\sum_{k=1}^{K} P(k \mid \boldsymbol{y}_t, \lambda)}$$

$$\hat{\Sigma}_n = \frac{\sum_{t=1}^{T}\sum_{k=1}^{K} P(k \mid \boldsymbol{y}_t, \lambda)[\Sigma_n(\boldsymbol{y}_t, k, \lambda) + \mu_n(\boldsymbol{y}_t, k, \lambda)\,\mu_n^{\mathrm{T}}(\boldsymbol{y}_t, k, \lambda)]}{\sum_{t=1}^{T}\sum_{k=1}^{K} P(k \mid y_t, \lambda)} - \hat{\mu}_n \hat{\mu}_n^{\mathrm{T}}$$

$$\hat{\boldsymbol{h}} = \sum_{t=1}^{T}\sum_{k=1}^{K} P(k \mid \boldsymbol{y}_t, \lambda)(((I-A)^{\mathrm{T}})^{-1}\Sigma_{x,k}^{-1}(\boldsymbol{y}_t - (I-A)\,\mu_{x,k} - A\boldsymbol{n} - B)$$

where

$$\mu_n(\boldsymbol{y}_t, k, \lambda) = \tilde{\Sigma}_n(\Sigma_n + \tilde{\Sigma}_n)^{-1}\,\mu_n + \Sigma_n(\Sigma_n + \tilde{\Sigma}_n)^{-1}\bar{\mu}_n$$
$$\Sigma_n(\boldsymbol{y}_t, k, \lambda) = [\Sigma_n^{-1} + \tilde{\Sigma}_n^{-1}]^{-1}$$

where

$$\bar{\mu}_n = (A^{\mathrm{T}})^{-1}(\boldsymbol{y}_t - (I-A)\,\mu_x - (I-A)\boldsymbol{h} - B]$$
$$\tilde{\Sigma}_n = (A^{\mathrm{T}})^{-1}(I-A)^{\mathrm{T}}\Sigma_{x,k}(I-A)(A^{\mathrm{T}})^{-1}$$

3.2 The estimation of non-stationary noise and channel

Now, the compensation is applied in cepstral domain. So the environment model is derived from (1):

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{h} + C\log(I + \mathrm{e}^{C^{\mathrm{T}}(\boldsymbol{n}-\boldsymbol{x}-\boldsymbol{h})}) \tag{7}$$

where C is the discrete cosine transform matrix.

Recursive noise parameter estimation is a solution to the recursive EM optimization problem[7~9].

$$\boldsymbol{n}_{t+1} = \arg\max_{n} Q_{t+1}(\boldsymbol{n})$$

The objective function $Q_{t+1}(\boldsymbol{n})$ above is the conditional expectation.

$$Q_{t+1}(\boldsymbol{n}) = E[\ln p(\boldsymbol{y}_1^{t+1}, \boldsymbol{K}_1^{t+1} \mid \boldsymbol{n}) \mid, \boldsymbol{y}_1^{t+1}, \boldsymbol{n}_1^{t}]$$

where $K_1^{t+1} = k_1, k_2, \cdots, k_{t+1}$ is the sequence of (hidden) mixture components in the clean speech model up to time $t+1$. By following the way similar to what is used in [9], we have

$$Q_{t+1}(\boldsymbol{n}) = \varepsilon \cdot Q_t(\boldsymbol{n}_t) - R_{t+1}(\boldsymbol{n}_{t+1}) \tag{8}$$

where

$$R_{t+1} = \sum_{k=1}^{K} \gamma_{t+1}(k)[\boldsymbol{y}_{t+1} - \mu_k^y(n_{t+1})]^{\mathrm{T}}(\Sigma_k^y)^{-1}[\boldsymbol{y}_{t+1} - \mu_k^y(n_{t+1})]$$

The forgetting factor ε is based on a tradeoff between the strength of noise tracking ability and the reliability of noise estimate. The occupancy probability $\gamma_\tau(k)$ is computed using Bayes rule in the E-step.

We can prove that $Q_{t+1}(n_{t+1})$ in recursive (7) is maximized via the following recursive form of noise parameter updating:

$$\boldsymbol{n}_{t+1} = \boldsymbol{n}_t + D_{t+1}^{-1} s_{t+1} \tag{9}$$

where

$$s_{t+1} = \frac{\partial R_{t+1}}{\partial n}\Big|_{n=n_t} = \sum_{k=1}^{K} \gamma_{t+1} A^{\mathrm{T}}(\Sigma_k^y)^{-1}(\boldsymbol{y}_{t+1} - \mu_k^y(\boldsymbol{n}_{t+1}))$$
$$D_{t+1} = \frac{\partial^2 Q_{t+1}}{\partial^2 \boldsymbol{n}}\Big|_{\boldsymbol{n}=\boldsymbol{n}_t} = -\sum_{\tau=1}^{t+1} \varepsilon^{t+1-\tau} \sum_{k=1}^{K} \gamma_\tau(k)(I-A)^{\mathrm{T}} \cdot (\Sigma_k^y)^{-1}(I-A)$$

In a similar way, we can get the formulas of the channel parameter $\boldsymbol{h}$. The practical algorithm execution steps in detail are described in [10].

4 Experimental results

4.1 Experiment setting

To evaluate the effectiveness of the proposed algorithm, we perform a series of experiments on telephone-quality and artificially contaminated speaker-independent (SI) Mandarin speech recognition.

In order to obtain telephone quality speech materials for the acoustic model training, we utilize the Mandarin 863 speech database developed by Chinese National "863" Program for LVCSR. It contains about 70 hours' speech. The speech data are 16kHz sampled and 16bit linearly quantized. We disposed the database with three methods: 1) resample the database to 8kHz with u-law quantization; 2) pass the database through the real PSTN network by Dialogic telephone cards plugged in PCs; 3) pass the database through GSM full rate (GSM FR 06. 10) coder and decoder. All these three transcoded databases are used as training data for the acoustic model.

The acoustic features consist of energy, pitch, 12 mel-cepstral with delta and delta-delta features. The vocabulary of this task consists of more than 40K words. Tri-gram statistics are used for language modeling.

There are three test sets in our experiments named as TELTEST, GSMTEST, SIMUTEST, respectively. TELTEST is gathered through the PSTN network, GSMTEST is gathered through GSM-FR codec and SIMUTEST is derived from artificial exhibition contamination. Every subset contains about 240 continuous Mandarin sentences from 4 different speakers.

The compensation of noise speech is based on MMSE. Once the parameters of the distribution of the noise speech (y) are computed, an MMSE estimate is used to calculate the clean speech from the observed noisy speech:

$$\hat{x}_{\mathrm{MMSE}} = E(x \mid y) = \int x p(x \mid y) \mathrm{d}x = y - \sum_{k=0}^{K-1} P(k \mid y)(\mu_{y,k} - \mu_{x,k})$$

We also can use an alternative approach to correct means and variances of HMM[10] instead of performing the MMSE estimate of clean speech.

4.2 Comparison of the actual function and our approximation

A comparative plot of the actual function and our approximation is shown in Fig. 3. As can be seen from this figure, the approximation can not be distinguished from the actual function.

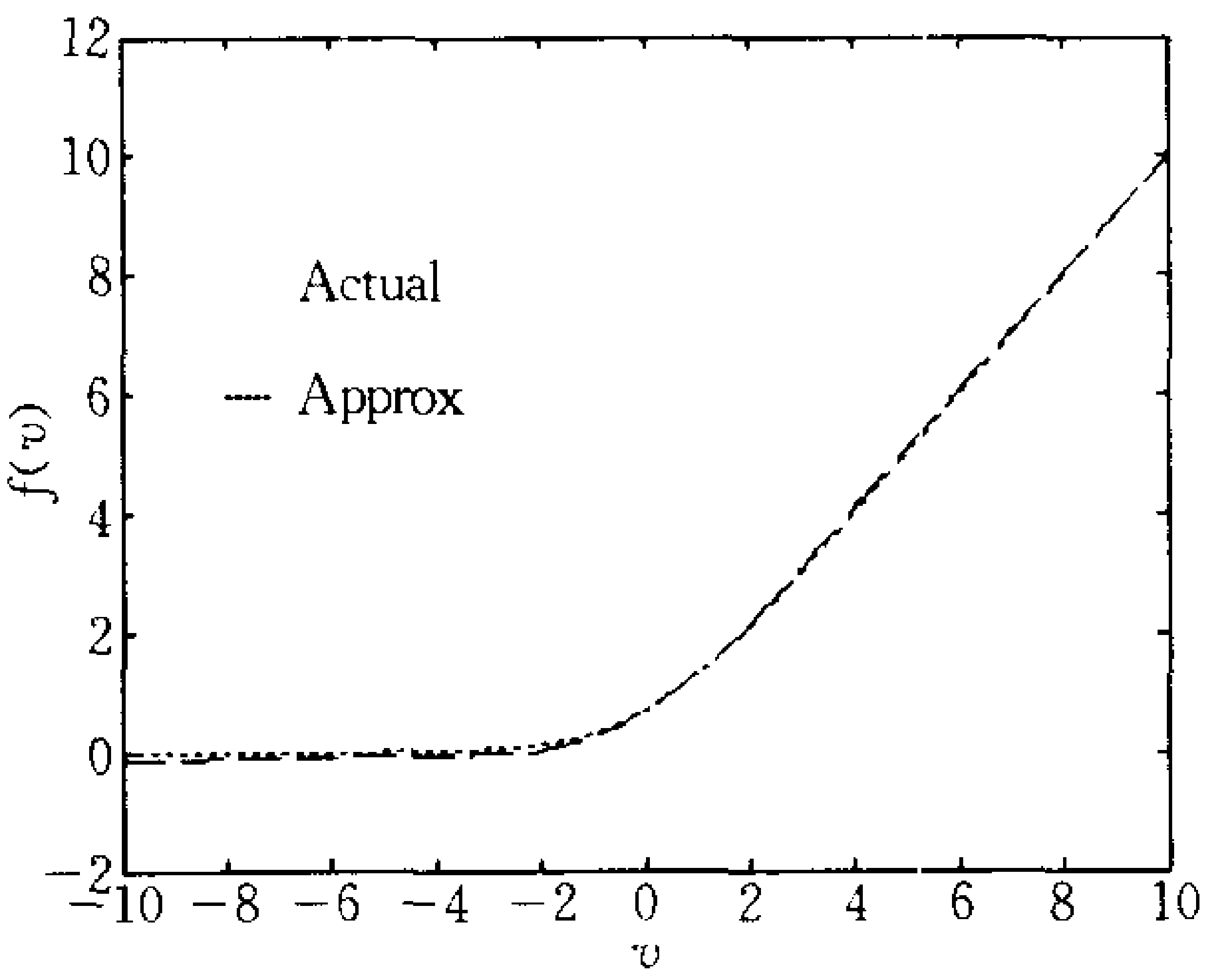

Fig. 3 Comparative plot of the environment function and the approximation function

4.3 Comparison of VPP and other methods

We compare our VPP algorithm with other three widely used compensation algorithms.

1) Long term CMS (Cepstral Mean Subtraction over the duration of a whole Mandarin sentence).

2) SBR (Signal Bias Removal)[11].

3) First order VTS[3].

The performance is measured by character error rate (CER) using NIST's SCLITE.

In Table 1, we present the results from the experiment using speaker-independent Mandarin speech recognition in telecommunication environment. VPP1 denotes the compensation applied to the batch EM, and VPP2 to the recursive EM.

Table 1 Results on telephone-quality speaker-independent Mandarin speech recognition

	Base-Line	CMS	SBR	VTS	VPP1	VPP2
TEL TEST	76. 4	78. 4	77. 1	78. 3	79. 2	80. 7
GSM TEST	75. 9	76. 3	78. 7	79. 8	80. 4	81. 1

From Table 1, VPP is the most effective one among these four compensation techniques. Compared with the baseline, VPP decreases the CER about 18% for TELTEST and 21. 5% for GSMTEST. Especially for GSMTEST, where GSM channel is more nonlinear than fixed lines, VPP suits it best.

Fig. 4 shows performance of VPP algorithm at several SNR's of SIMUTEST. We find that VPP obtains significantly lower CERs.

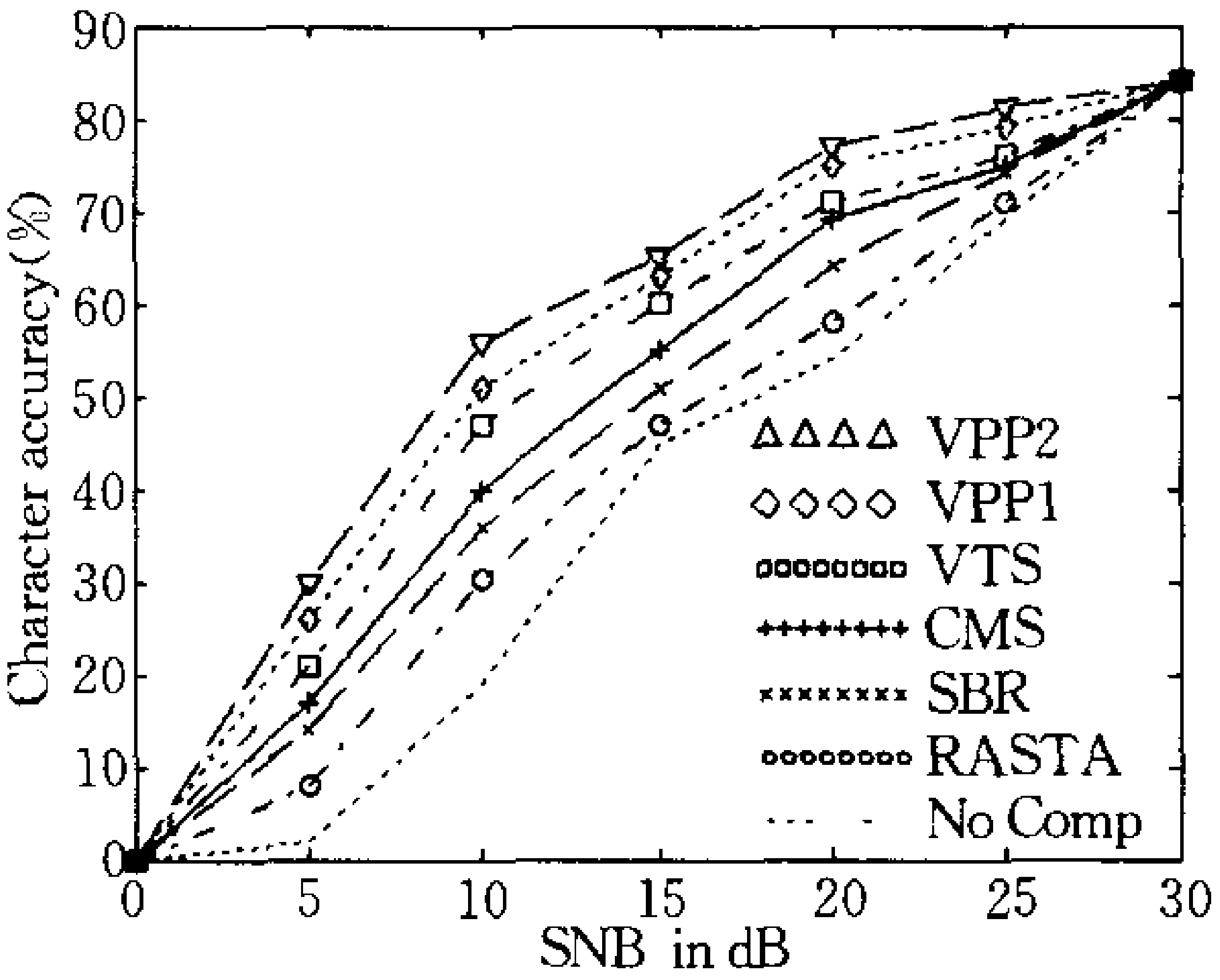

Fig. 4 Performance of VPP algorithm at several SNR's

In the experiment of VPP1, the covariance matrices of the clean speech, the noisy speech and the additive noises are assumed to be diagonal in order to reduce the computational complexity. In the future, we will do some experiments on using the full matrices.

5 Conclusions

In this paper, we have presented a novel approximation—based method to compensate for the effects of linear filter and background noise given the HMM parameters of clean speech. The algorithm is successfully applied to the batch EM and the recursive EM. The experimental results show that the approximation of environment function using the VPP algorithm is more precise than VTS. In the telephone network, the algorithm presented here provides significant improvement over the previous work in data compensation. Compared with other compensation approaches, VPP shows better adaptability to variations introduced by channel and noise, and obtains about 18% character error rate decrease in Mandarin telephone speech recognition.

References

1 Chien J T, Wang H C, Lee L M. Estimation of channel bias for telephone speech recognition. In: Proceedings of the 4th International Conference on Spoken Language Processing, USA, 1996. 1840~1843

2 Acero A. Acoustical and environmental robustness in automatic speech recognition[Ph. D. dissertation]. Department of Electrical and Computer Engineering, Carnegie Mellon University, 1990. 72~85

3 Moreno P J. Speech recognition in noisy environments[Ph. D. dissertation]. Department of Electrical and Computer Engineering, Carnegie Mellon University, 1996. 79~96

4 Moreno P J, Raj B, Gouvea E B, Stern R M. Multivariate Gaussian based cepstral normalization for robust speech recognition. In: IEEE Conference on Acoustics, Speech and Signal Processing, Detroit, 1995. 137~140

5 Raj, Gouvêa E, Stern R M. Vector polynomial approximations for robust speech recognition. In: Proceedings of the ESCA Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels, France, 1997

6 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): 39～49

7 Kim N S. Nonstationary env ronment compensation based on sequential estimation. *IEEE Signal Processing Letters*, 1998, **5**(3): 57～60

8 Krishnamurthy V, Morre J B. Online estimation of hidden Markov model parameters based on the Kullback-Leibler information feature. *IEEE Transactions on Signal Processing*, 1993, **41**(8): 2557～2573

9 Deng L, Droppo J, Acero A. Recursive noise estimation using iterative stochastic approximation for stereo-based robust speech recognition. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, Italy, 2001

10 Acero A, Deng L, Kristjansson T, Zhang J. Hmm adaptation using vector Taylor series for noisy speech recognition. In: Proceedings of the 6th International Conference on Spoken Language Processing, Beijing, 2000, III: 869～872

11 Rahim M G, Juang Biing-Hwang. Signal bias removal by maximum likelihood estimation for robust telephone speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1996, **4**(1): 19～30

HAN Zhao-Bing Received his bachelor degree from Tsinghua University, and currently is a Ph. D. candidate at Institute of Automation, Chinese Academy of Sciences. His research interests include telephone speech recognition.

ZHANG Shu-Wu Received his Ph. D. degree from NLPR, Chinese Academy of Sciences in 1997. He was an invited researcher in ATR Spoken Language Translation Laboratories, Japan during April, 1998 to May, 2002. Currently, he is an associate professor of Institute of Automation, Chinese Academy of Sciences. His research interests include multilingual speech recognition, natural language processing, and web-based multimedia technologies.

XU Bo Received his bachelor degree from Zhejiang University in 1988, master and Ph. D. degrees from Institute of Automation, Chinese Academy of Sciences, in 1992 and 1997, respectively. His research interests include multilingual speech recognition, natural language processing, and intelligent information processing.

HUANG Tai-Yi Professor. His research interests include speech communication, natural language processing, and intelligent information processing.

电话语音识别中统一的加性噪声和卷积噪声补偿算法

韩兆兵　张树武　徐　波　黄泰翼

(中国科学院自动化研究所模式识别国家重点实验室　北京　100080)

(E-mail: {zbhan, swzhang, xubo, huang}@hitic.ia.ac.cn)

摘　要　为了统一地补偿电话语音受加性噪声和卷积通道响应的影响，本文提出了矢量分段多项式近似(VPP)算法. 并把此算法成功地应用到稳态噪声和非稳态噪声环境. 对于稳态噪声环境，在log谱域采用Batch EM(B EM)方法；对于非稳态噪声环境，在倒谱域采用递归EM(R EM)方法. 这两种方法都是基于最小均方误差估计(MMSE)准则的特征补偿. 实验结果表明，受背景噪声和电话通道(包括固定电话和GSM)影响的大词汇量连续语音识别应用此算法误识率可以降低约18%.

关键词　语音识别，分段多项式近似，环境补偿，递归EM算法

中图分类号　TP391.41