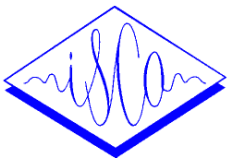

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

THE EFFICIENT PMC FOR ROBUST SPEECH RECOGNITION IN NOISY ENVIRONMENTS

Cailian MIAO
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing
clmiao@nlpr.ia.ac.cn

Yangsheng WANG
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing
wys@nlpr.ia.ac.cn

ABSTRACT

The environment adaptive methods play an important part in improving the robustness of automatic speech recognition. In this paper, PMC is reviewed and improved to achieve the better performance. The experiments have been done based on the Cambridge's HTK toolkit to implement the continuous Mandarin digit recognition in noisy environments

1. INTRODUCTION

Many researches have been done on robust speech recognition. The researches are summarized as the environmental compensation and the environmental adaptation.

The former concentrates on signal enhancement and noise level reduction. There is CMN, RASTA, spectral subtraction, Wiener filter and their extensions.

The latter transforms the clean speech models to adapt to noisy environments, which can be classified as the model transference and model transformation.

The representative approach of transference is PMC approach. PMC combines the clean speech HMM models with estimated noise model together and re-constructs the noisy speech models artificially. The PMC originated from HMM composition approach proposed by Varga & Moore (1990), was defined as PMC by Gales & Young (1993). The DPMC proposed by Gales & Young (1995), tied-state PMC by Komori & Kosaka (1997), weighted PMC by Tai-Hwei Hwang & Hsiao-Chuan Wang (2000), PCA-PMC by Ruhi Sarikaya (2000), VTS approximate solution proposed by Acero (2000) reduce the computational cost and advance PMC in some extent. The representative approach of transformation is MLLR.

PMC can model the noise independently and produce the new noisy speech models. This approach can get around additive noise as well as convolutional noise. In PMC solution, the estimation of convolutional noise model remains an open problem, meantime it is hard to get a closed-form expression of the dynamic noisy speech model parameters. In this paper, these problems are overcome in some extent.

In our experiment, Cambridge's HTK toolkit 3.0 was used as test platform with suitable modification to implement the speaker-dependent continuous Mandarin digit recognition. The training data were collected in clean office environment while the test data included the noisy speech collected in high background noise environment and the clean speech contaminated artificially by white Gaussian noise at different SNR levels.

The paper is organized as follow: In the Section 2, the PMC is briefly reviewed, the method of estimation of convolutional noise model is presented, and some modifications are introduced. In Section 3, an overview of the experimental condition is given and the experiment results are provided to show the performance of the proposed schemes. In Section 4, the future work is suggested. Finally the conclusion is drawn.

2. THE PMC

2.1 The review of PMC

The PMC (parallel model combination) provides an effective adaptation method to obtain the models of noisy speech without retraining database. First the combinations of mean and variance vector are performed in cepstral domain for convolutional noise and speech models, then the combinations are performed in spectral domain for additive noise and speech models. The whole procedure can be equated as following:

$$\lambda_O^{cep} = DCT\left\{\log\left[g \cdot \exp\left(IDCT\left(\lambda_S^{cep} + \lambda_H^{cep}\right)\right) + \exp\left(IDCT\left(\lambda_N^{cep}\right)\right)\right]\right\} \quad (1)$$

Where λ is model parameter, it's often described as HMM. Subscript O, S, N and H are the parameter of combined data, clean speech data, additive noise and convolutional noise, respectively. The superscript ***cep*** means a parameter is in cepstral domain. Similarly superscript **log** and ***lin*** are parameters in log-spectral domain and spectral domain, respectively. The ***DCT*** and ***IDCT*** are the discrete cosine transformation and the inverse transformation. g is a gain matching term introduced to account for level differences between the clean and noisy speech.

With the approach above, the statistics models of additive and convolutional noise are incorporated together to obtain the new noisy speech models. The mappings of model parameters between cepstral and spectral domain are termed as lognormal approximation, log-add approximation or data-driven PMC.

2.2 Estimation of Convolutional Noise

The model of additive noise is easily collected during a period of silence in ambient environments. It is unpractical and hard to get the model of the convolutional noise in real situation. Reverberant, microphone equipments and channel usually cause convolutional noise. In this paper, a little amount of adaptation data is required to get the statistical model parameter of convolutional noise $\boldsymbol{H}$.

The ML estimation is often used as the optimization criterion for estimating the convolutional noise $\boldsymbol{H}$. In our paper, the ML estimation is performed using the EM algorithm with two steps. In the first step, with the known noise model $\boldsymbol{\lambda}_N$, the integration models $\boldsymbol{\lambda}_{SH}^{lin}$ that incorporate clean speech models with convolutional noise can be derived in spectral domain with ML estimation as following:

$$\hat{\boldsymbol{\lambda}}_{SH}^{lin} = \arg\max P\left(\boldsymbol{O}^{lin} \middle| \boldsymbol{\lambda}_{SH}^{lin}, \boldsymbol{\lambda}_N^{lin}\right) \tag{2}$$

The probability density function of integration models $\boldsymbol{\lambda}_{SH}^{lin}$ is assumed to be a mixture of Gaussian densities. In order to estimate the integration models $\boldsymbol{\lambda}_{SH}^{lin}$, the auxiliary function is defined as following:

$$Q(\boldsymbol{\lambda}_{SH}^{lin}, \hat{\boldsymbol{\lambda}}_{SH}^{lin}) = \sum_{n=1}^{W_n} \sum_{\theta \in \Theta^n} \sum_{K \in \Omega_m^n} \frac{P\left(\boldsymbol{O}^{lin}, \theta, K \middle| \boldsymbol{\lambda}_{SH}^{lin}, \boldsymbol{\lambda}_N^{lin}\right)}{P\left(\boldsymbol{O}^{lin} \middle| \boldsymbol{\lambda}_{SH}^{lin}, \boldsymbol{\lambda}_N^{lin}\right)} \log P\left(\boldsymbol{O}^{lin}, \theta, K \middle| \hat{\boldsymbol{\lambda}}_{SH}^{lin}, \boldsymbol{\lambda}_N^{lin}\right)$$

$$= \sum_{n=1}^{W_n} \sum_{t=1}^{T_n} \gamma_{jk}^{lin}(t) \log b_{jk}^{lin}\left(\boldsymbol{O}^{lin}(t)\right) \tag{3}$$

Where W_n is number of adaptation sentence, n is the frame index of current sentence, $\theta \in \Theta$ is the all possible state sequences, $K \in \Omega_m$ is the set of all possible mixture component branch sequences. $b_{jk}^{lin}(\boldsymbol{O}(t))$ is the Gaussian PDF in mixture component k of state j at time t given the observation $\boldsymbol{O}$, and $\gamma_{jk}^{lin}(t)$ is corresponding occupation probability defined as following:

$$\gamma_{jk}^{lin}(t) = \frac{1}{P\left(\boldsymbol{O} \middle| \boldsymbol{\lambda}_{SH}^{lin}\right)} \sum_{\theta \in \Theta^n} \sum_{K \in \Omega_m^n} P\left(\boldsymbol{O}^{lin}, \theta_t = j, K_t = k \middle| \boldsymbol{\lambda}_{SH}^{lin}, \boldsymbol{\lambda}_N^{lin}\right) \tag{4}$$

Taking partifical derivative of Equ. (4) with respect to the mean $\boldsymbol{\mu}_{SH}^{lin}$ and variance $\boldsymbol{\Sigma}_{SH}^{lin}$ parameter, setting the equations to zero, we can get the solution of the integration models:

$$\frac{\partial Q(\boldsymbol{\lambda}_{SH}^{lin}, \hat{\boldsymbol{\lambda}}_{SH}^{lin})}{\partial \boldsymbol{\mu}_{SH}^{lin}} = \boldsymbol{0} \qquad \frac{\partial Q(\boldsymbol{\lambda}_{SH}^{lin}, \hat{\boldsymbol{\lambda}}_{SH}^{lin})}{\partial \boldsymbol{\Sigma}_{SH}^{lin}} = \boldsymbol{0} \tag{5}$$

Given the models of clean speech $\boldsymbol{\lambda}_S$, the second step is separating the model of convolutional noise from the integration models $\boldsymbol{\lambda}_{SH}$. First, all parameters are mapped to cepstral domain. Then the auxiliary function is defined:

$$Q(\boldsymbol{\lambda}_H^{cep}, \hat{\boldsymbol{\lambda}}_H^{cep}) = \sum_{n=1}^{W_n} \sum_{\theta \in \Theta^n} \sum_{K \in \Omega_m^n} \frac{P\left(\boldsymbol{O}', \theta, K \middle| \boldsymbol{\lambda}_H^{cep}, \boldsymbol{\lambda}_S^{cep}\right)}{P\left(\boldsymbol{O}' \middle| \boldsymbol{\lambda}_H^{cep}, \boldsymbol{\lambda}_S^{cep}\right)} \log P\left(\boldsymbol{O}', \theta, K \middle| \hat{\boldsymbol{\lambda}}_H^{cep}, \boldsymbol{\lambda}_S^{cep}\right) \tag{6}$$

Where $\boldsymbol{O}'$ is the feature vectors presented by the mean parameter of integration model as following:

$$\sum_{n=1}^{W_n} \sum_{t=1}^{T_n} \gamma_{jk}^{l}(t) \boldsymbol{O}'(t) = \left(\sum_{t=1}^{T_n} \gamma_{jk}(t) \right) \hat{\boldsymbol{\mu}}_{SH}^{cep} \tag{7}$$

With the similar EM algorithm, the model $\boldsymbol{\lambda}_H^{cep}$ is estimated.

Now we can combine the model of speech λ_S ,noise λ_N and λ_H together to get noisy speech models. With such updated models, we repeat the estimation procedure above for 1~4 times to get more optimal models until the result converges.

With the limited adaptation data, only a global model of the convolutional noise λ_H can be obtained. When data increases, the individual models λ_H can be generated corresponding to different speech models.

2.3 Mapping scheme modification

The closed-form formulations of mapping the model parameters between cepstral and spectral domain based on the lognormal assumption cost high computation. It becomes more critical in iterative processing. A modified log-add scheme is proposed to solve this problem.

In conventional log-add scheme the spectral mean vector of noisy speech models is approximated by:

$$\boldsymbol{\mu}_O^{lin} \cong g \cdot \boldsymbol{\mu}_S^{lin} \cdot \boldsymbol{\mu}_H^{lin} + \boldsymbol{\mu}_N^{lin} \tag{8}$$

Where $\boldsymbol{\mu}_S^{lin} = \exp\left(\boldsymbol{\mu}_S^{\log}\right)$ $\boldsymbol{\mu}_H^{lin} = \exp\left(\boldsymbol{\mu}_H^{\log}\right)$ $\boldsymbol{\mu}_N^{lin} = \exp\left(\boldsymbol{\mu}_N^{\log}\right)$

Similarly, the spectral mean vector of integration model is:

$$\boldsymbol{\mu}_{SH}^{lin} = \boldsymbol{\mu}_S^{lin} \cdot \boldsymbol{\mu}_H^{lin} \tag{9}$$

Since the variance is not processed, the variance vector of noisy speech models is retained from the original speech models. The log-add scheme is simple, however the performance is limited.

To solve this problem, the scalars $\boldsymbol{\alpha}_{SH}$ and $\boldsymbol{\alpha}_N$ are introduced to balance the influence of variance vector between noise and integration speech models:

$$\boldsymbol{\Sigma}_O^{cep} \cong \frac{\boldsymbol{\alpha}_{SH}}{\boldsymbol{\alpha}_{SH} + \boldsymbol{\alpha}_N} \boldsymbol{\Sigma}_{SH}^{cep} + \frac{\boldsymbol{\alpha}_N}{\boldsymbol{\alpha}_{SH} + \boldsymbol{\alpha}_N} \boldsymbol{\Sigma}_N^{cep} \tag{10}$$

So the variance vector of noisy speech models can be directly obtained from Equ. (10).

It is assumed that all covariance matrixes are diagonal. The scalars $\boldsymbol{\alpha}_{SH}$ and $\boldsymbol{\alpha}_N$ is shown as following:

$$\boldsymbol{\alpha}_{SH} = \frac{1}{\boldsymbol{D}} \sum_{d=1}^{\boldsymbol{D}} \boldsymbol{\xi}_d^2 \qquad \boldsymbol{\alpha}_{SH} = \frac{1}{\boldsymbol{D}} \sum_{d=1}^{\boldsymbol{D}} \bar{\boldsymbol{\xi}}_d^2 \tag{11}$$

Where $\boldsymbol{\xi}_d = \frac{g\boldsymbol{\mu}_{SH,d}^{lin}}{\boldsymbol{\mu}_{O,d}^{lin}}$ $\bar{\boldsymbol{\xi}}_d = \frac{\boldsymbol{\mu}_{N,d}^{lin}}{\boldsymbol{\mu}_{O,d}^{lin}}$, $\boldsymbol{D}$ is the dimension of the vector.

There are some benefits to introduce the scalars: it contacts the variance of noise and speech models together. When signal gain is in high lever, the speech models exert predominant role, vice versa. The variance vector is gotten around directly in cepstral domain. Without huge computation cost, modified log-add scheme improve the recognition accuracy.

2.4 The dynamic parameter

Even though some approximate approaches were used, the lognormal approximation can't handle the dynamic parameter exactly. The closed-form formulation to describe the dynamic parameter in log-add schemes can be written as following:

$$\Delta\mu_O^{cep} \approx DCT\left[\left(\frac{g\cdot\mu_S^{lin}}{g\cdot\mu_{SH}^{lin}+\mu_N^{lin}}\right)\cdot\Delta\mu_S^{\log}+\left(\frac{\mu_N^{lin}}{g\cdot\mu_{SH}^{lin}+\mu_N^{lin}}\right)\cdot\Delta\mu_N^{\log}\right] \quad (12)$$

$$\Delta^2\mu_O^{cep} \approx DCT\left[\left(\frac{g\cdot\mu_S^{lin}}{g\cdot\mu_{SH}^{lin}+\mu_N^{lin}}\right)\cdot\Delta^2\mu_S^{\log}+\left(\frac{\mu_N^{lin}}{g\mu_{SH}^{lin}+\mu_N^{lin}}\right)\cdot\Delta^2\mu_N^{\log}\right.$$
$$\left.+\left(\frac{g\mu_S^{lin}\cdot\mu_N^{lin}}{\left(g\mu_{SH}^{lin}+\mu_N^{lin}\right)^2}\right)\left(\Delta\Sigma_S^{\log}+\left(\Delta\mu_S^{\log}\right)^2+\Delta\Sigma_N^{\log}\right)\right] \quad (13)$$

$$\Delta\Sigma_O^{cep} \cong \frac{\alpha_{SH}}{\alpha_{SH}+\alpha_N}\Delta\Sigma_{SH}^{cep}+\frac{\alpha_N}{\alpha_{SH}+\alpha_N}\Delta\Sigma_N^{cep} \quad (14)$$

$$\Delta^2\Sigma_O^{cep} \cong \frac{\alpha_{SH}}{\alpha_{SH}+\alpha_N}\Delta^2\Sigma_{SH}^{cep}+\frac{\alpha_N}{\alpha_{SH}+\alpha_N}\Delta^2\Sigma_N^{cep} \quad (15)$$

The equations can be applied not only for simple form but also the linear regression dynamical coefficients. In Equ. (12) and (13) , the dynamic parameter of mean vector μ_H is ignored. The scalars are also used in dynamical variance matrixes.

Additionally, the feature vectors are often truncated to the lower orders cepstral parameters, so that the dimension of the cepstrum vector is much smaller than the number of filter banks. This means that it is not possible to perform the inverse DCT transformation to go back to log-spectral domain. In order to get around this problem, means and variances are often zero padded to the required number to allow the inverse DCT to be achieved. Alternative approach is presented in this paper, the DCT matrix is forced be square by truncation preliminarily. The experiment shows there is no any performance degradation.

3. EXPERIMENTS

In this paper, some experiences were carried out to evaluate the proposed algorithms. HTK toolkit was used as test platform with some modification to implement the speaker-dependent continuous Mandarin digit recognition.

Continuous mandarin digit recognition is a well typical example to show the performance of our study, it is more difficult than English digit recognition. Considering such reason and the application, the system was designed to recognize the continuous mandarin digits.

3.1 Experimental condition

The original clean speech data were recorded using a normal close-talking microphone in an office room. The average SNR is about 25B. All speech data were sampled at 8kHz and quantified in 16bit. 1000 sentences of random digit numbers (containing 3~8 numbers per sentence) read in naturally continuously speaking way were used as training data while another 200 sentences as test data.

The pre-emphasized feature vector of speech data is described by the MFCC including 39 vectors, (13 MFCC, the delta and acceleration difference components obtained by linear regression filtering) windowed in 25 ms Hamming in 10 ms step with 26 filter banks. Each phone model has 5 states with single mixture Gaussian distribution.

The test speech data included the artificial corrupted speech by adding the white noise to the clean speech at different SNR levels, and the real noisy speech by recording the speech using another desktop microphone in high background noise. The real noisy data contained 500 sentences with SNR ranging from 4dB to 15 dB. The average SNR was less than 10 dB. They were divided into 13 groups independently.

3.2 PMC approach

At first, the PMC is performed in artificial additive noise test. The result is summarized in Table.1. In matched condition, the accuracy of baseline system is 85.96%, and the performances tend to be significantly degraded in low SNR, especially be 20.66% in –5dB.

SNR (dB)	Baseline (%)	PMC1 (%)	PMC2 (%)
-5	20.66	20.66	45.11
0	54.89	54.89	71.29
5	75.55	75.55	79.92
10	81.55	81.55	82.02
15	86.12	86.12	85.65
Clean	86.96	86.96	—

Table 1: *The accuracy of recognition of Baseline system, PMC1 approaches, where PMC1 indicates the result of truncated DCT matrix without combining noise mode, PMC2 combines the speech and noise models*

In real noisy environment, it is easy to collect the background noise. The 2-step EM-estimation is used to get the model of the convolutional noise. The energy parameter is used to adjust the speech and noise spectral levels.

Figure 1. (a) and (b) show the accuracy of recognition with respect to the iteration times and adaptation data. From Figure 1, it is concluded that the 3 iterative times is enough and with only 10 adaptation data, the PMC can work well.

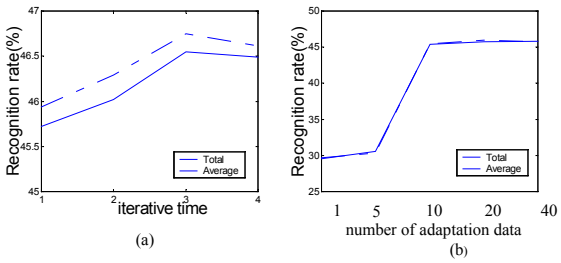

Figure 1: The accuracy of recognition in PMC in real noise environment, where the 'Total' indicates accuracy of whole 500 sentences; the 'Average' is the average accuracy of the 13 groups.

Table 2 shows the accuracy of baseline and kinds of PMC approaches. The PMC3 uses the lognormal approximation; the PMC4 uses the conventional log-add approach with unaltered clean speech model variance vector; PMC5 uses the log-add approach with variance vector of the estimated integration model λ_{SH} , PMC6 uses the log-add with variance adaptation

approach shown in Equ (10). Only the 13-order static parameters are used.

Scheme	Total	Average
Baseline (%)	19.41	19.25
PMC-3 (%)	45.72	45.98
PMC-4 (%)	29.09	28.80
PMC-5 (%)	42.54	42.73
PMC-6 (%)	42.18	42.39

Table 2: The accuracy of recognition in Baseline system and kinds of PMC approaches for the first iterative time with 20 adaptation data.

At last, Table 3 shows the result using the log-add PMC approach with the variance adaptation (PMC5), when the dynamic parameter is added.

Scheme	Total	Average
Baseline (%)- Δ , Δ^2	31.33	30.57
PMC-2 (%)- Δ , Δ^2	32.33	31.68
PMC-6 (%)- Δ , Δ^2	43.42	43.91

Table 2: The accuracy of recognition in Baseline system, the PMC only combine the additive noise (PMC-2), and the PMC-6 approach with the dynamical parameter for the first iterative time with 20 adaptation data (PMC-5).

4. DISCUSSION

The experiences show that the PMC approach is an effective scheme in robust speech recognition. In real condition, the estimation of convolutional noise model and the introduction of dynamic can highly improve the performance.

The PMC can be improved and extended in a number of ways. One is to find a more accurate method to estimate the convolutional noise, the other is to incorporate dynamic parameter throughout the whole procedure, while in this paper, they are only be introduced at the combination step.

5. CONCLUSIONS

In this paper, the PMC approach is reviewed and further extended to make the performance better. It provides some useful information to improve the robustness of speech recognition.

6. REFERENCES

[1] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "Hmm Adaptation Using Vector Taylor Series For Noisy Speech Recognition", *Proceedings of ICSLP*, 2000.

[2] X.D. Huang, A. Acero, and H. Hon, "*Spoken Language Processing*", Prentice Hall, 2000

[3] R. Sarikaya, J.H.L. Hansen, "PCA-PMC: A Novel Use Of A Priori Knowledge For Fast Parallel Model Combination", *proceedings ICASSP*, vol. II, pp. 1113-1116, (SP-L6.3.1-4), Istanbul, Turkey, 2000

[4] S. -J. Doh, "Enhancements to Transformation-Based Speaker Adaptation: Principal Component and Inter-Class Maximum Likelihood Linear Regression", *Ph.D. Thesis*, ECE Department, CMU, 2000

[5] Young S. Et al, The HTKBook, http:// htk.eng.cam.ac.uk/, 2001

[6] Lawrence Rabiner, Biing-Hwang Juang, "Fundamentals of speech recognition", Prentice-Hall, 1996

[7] Volker Stahl, Alexander Fischer, Rolf Bippus, "Acoustic synthesis of training data for speech recognition in living room environment", *Proc. ICASSP'2001*, Salt Lake City, Utah, May 7-11,2001

[8] M.J.F. Gales "The Generation And Use Of Regression Class Trees For MLLR Adaptation", *Technical Report CUED/F-INFENG/TR.263,* Cambridge University Engineering Department, 1996

[9] T. Kristjansson, B. Frey, L. Deng and A. Acero, "Towards Non-Stationary Model-Based Noise Adaptation for Large Vocabulary Speech Recognition", *Proc-ICASSP'2001,* Salt Lake City, Utah, 2001

[10] P.C. Woodland, M.J.F. Gales & D.Pye, "Improving Environmental Robustness In Large Vocabulary Speech Recognition", *Proc. ICASSP'1996*, Atlanta, GA 1996.

[11] Tai-Hwei Hwang and Hsiao-Chuan Wang, "A Fast Algorithm For Parallel Model Combination For Noisy Speech Recognition", *Computer Speech And Language,* No.14, p81-100,2000

[12] B. Frey, L. Deng, A. Acero & T. Kristjansson, "ALGONQUIN: Iterating Laplace's Method to Remove Multiple Types of Acoustic Distortion for Robust Speech Recognition", *Proc. Eurospeec'2001*, Aalborg, Denmark, 2001

[13] Tetsuya Takiguchi, Satoshi Nakamura, Kiyohiro Shikano, "HMM-separation-Based Speech Recognition for a Distant Moving Speaker", *IEEE Transactions On Speech And Audio Processing*, Vol9, No.2 February, 2001

[14] Rahim,M,G. and B.H.Juangv, "Signal Bias Removal By Maximum Likelihood Estimation For Robust Telephone Speech Recognition", *IEEE Transactions On Speech And Audio Processing*, Vo4, No.1, 1996

[15] P.J. Moreno, "Speech Recognition in Noisy Environments", *Ph.D. thesis*, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213,1996

[16] M.J.F. Gales, "Model-Based Techniques for Noise Robust Speech Recognition", *Ph.D. thesis*, University of Cambridge, 1996

[17] A.P. Varga and R.K. Moore, "Hidden Markov Model Decomposition Of Speech And Noise", Proc. ICASSP'1990, pp. 845–848, 1990

[18] Ivandro Sanches "Noise-Compensated Hidden Markov Models", *IEEE Transactions On Speech And Audio Processing*, Vo8, No.5 September, 2000

[19] Tetsuya Takigughi, Satoshi Nakamura, Qiang Huo & Kiyohiro, "Model Adaptation Based On HMM Decomposition For Reverberant Speech Recognition", *proceedings ICASSP-1997*, pp. 827-830,1997