Improved Robustness for Speech Recognition Under Noisy Conditions Using Correlated Parallel Model Combination

Jeih-weih Hung, Jia-lin Shen and Lin-shan Lee
Institute of Information Science, Academia Sinica
Nankang, Taipei, Taiwan, Republic of China

Abstract

The parallel model combination (PMC) technique has been shown to achieve very good performance for speech recognition under noisy conditions. In this approach, the speech signal and the noise are assumed uncorrelated during modeling. In this paper, a new correlated PMC is proposed by properly estimating and modeling the nonzero correlation between the speech signal and the noise. Preliminary experimental results show that this correlated PMC can provide significant improvements over the original PMC in terms of both the model differences and the recognition accuracies. Error rate reduction on the order of 14% can be achieved.

1. Introduction

While practically used in real world applications, many speech recognition systems often can't perform as well as it could during the controlled training environment. This performance degradation is mainly due to the mismatch between the training and test conditions, in which the mismatch in noisy conditions is usually very important. In order to achieve robust speech recognition in various noise conditions, substantial research efforts have been made and many good algorithms have been proposed. Some of them used noise resistant speech feature parameters. Some of them tried to perform the compensation in either the speech feature domain or the speech model domain. For the speech model compensation methods, the Parallel Model Combination Method (PMC) [1, 2] has been one of the most popular approaches with its effectiveness to handle the additive noise well confirmed by many experiments and systems. With the PMC method the approximated noisy speech HMM's can be derived as long as the Noise HMM and the pre-trained clean speech HMM's are available. In this way the difficulties in collecting enormous noisy speech data to retrain the noisy speech HMM's can be avoided, and only a small amount of noise data are required to train the noise HMM. The key concept of PMC is that the noise is additive in the linear spectral domain. Therefore the cepstral-based parameters of the clean speech HMM's and the noise HMM must be transformed to the log-spectral domain or the linear spectral domain in order to perform the combination, and then inversely transformed back to the cepstral domain for normal recognition process.

Nevertheless, the linear combination of speech and additive noise in the linear spectral domain does not necessarily provide very good results in many cases as expected [3]. One possible reason for it, among many others, may be due to the possible correlation between speech and noise, which was not modeled very well in the original PMC method. Such correlation can be observed during the short-term analysis for feature extraction, and the nonzero noise mean could be a possible source of it [4], as will be clear later on. A good example phenomenon is that the PMC method usually doesn't operate as well while SNR becomes worse, in which the nonzero noise mean may become larger. As a result, it is expected that if such correlation between speech and noise can be adequately considered and well modeled, the accuracy of the resulting HMM's may be improved. Such an approach is presented in this paper, and the initial experimental results indicate that this may be a correct direction.

The remainder of the paper is organized into 4 sections. The PMC method is briefly summarized for further development purposes in section 2. The correlation mentioned above is then discussed and modeled in the basic PMC method in section 3. Section 4 then presents some preliminary experimental results using this correlated PMC method.

2. Parallel Model Combination

The PMC method generates noisy-speech-like models by combining clean speech HMM's and a noise HMM. The HMM cepstral parameters must be mapped to the linear spectral domain for combination. The detailed procedures can be summarized as the following steps, where $\boldsymbol{\mu}^c$ and $\boldsymbol{\Sigma}^c$ are the clean speech mean and variance respectively in the cepstral domain, $\boldsymbol{\mu}^l$ and $\boldsymbol{\Sigma}^l$ the log spectral domain, and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ the linear spectral domain. Where the noise distributions need to be indicated a "$\hat{}$" is used and for the estimated corrupted speech distributions a "$\tilde{}$", thus for instance, $\hat{\boldsymbol{\mu}}^c$ and $\tilde{\boldsymbol{\mu}}^c$ are the noise and estimated noisy speech cepstral means respectively. The factor *g* is a gain matching term dependent on SNR.

1. Inverse DCT transformation

$$\boldsymbol{\mu}^l = \mathbf{C}^{-1}\boldsymbol{\mu}^c \tag{1}$$

$$\boldsymbol{\Sigma}^l = \mathbf{C}^{-1}\boldsymbol{\Sigma}^c(\mathbf{C}^{-1})^T \tag{2}$$

where **C** is the matrix for DCT.

2. Exponential transformation

$$\mu_i = \exp\left(\mu_i^l + \Sigma_{ii}^l/2\right) \quad (3)$$

$$\Sigma_{ij} = \mu_i \mu_j \left[\exp\left(\Sigma_{ij}^l\right) - 1\right] \quad (4)$$

3. Composition

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\mu} + g\tilde{\boldsymbol{\mu}} \quad (5)$$

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} + g^2\tilde{\boldsymbol{\Sigma}} \quad (6)$$

4. Logarithm transformation

$$\hat{\mu}_i^l = \log\left(\hat{\mu}_i\right) - \frac{1}{2}\log\left(\frac{\hat{\Sigma}_{ii}}{\hat{\mu}_i^2} + 1\right) \quad (7)$$

$$\hat{\Sigma}_{ij}^l = \log\left(\frac{\hat{\Sigma}_{ij}}{\hat{\mu}_i \hat{\mu}_j} + 1\right) \quad (8)$$

5. DCT transformation

$$\hat{\boldsymbol{\mu}}^c = \mathbf{C}\hat{\boldsymbol{\mu}}^l \quad (9)$$

$$\hat{\boldsymbol{\Sigma}}^c = \mathbf{C}\hat{\boldsymbol{\Sigma}}^l\mathbf{C}^T \quad (10)$$

3. Correlated PMC

Although under most noisy environments the PMC method can provide significant improvements in recognition rates as compared with the HMM's trained with clean speech, very often there still exists a performance gap between PMC and the HMM's trained by noisy speech under matched conditions, and usually this gap becomes wider when SNR becomes worse. There can be many reasons for this. The inaccuracy in the log-normal approximation in the logarithm transformation process could be one, while the correlation between the speech signal and noise as pointed out here could be another. This will be discussed below.

Let S, N and X be the spectra of the clean speech, the noise and the resulting noisy speech signal, respectively,

$$X = S + N \quad (11)$$

During the feature extraction process, the signal plus noise was filtered by a set of Mel-scale filters in parallel. At the output of filter *j*, the power of the noisy speech signal for a given signal frame, $|x_j|^2$, is given by

$$|X_j|^2 = |S_j + N_j|^2 = |S_j|^2 + |N_j|^2 + 2\,\mathrm{Re}\{S_j N_j^*\} \quad (12)$$

where S_j and N_j are the corresponding filter output for clean speech and noise respectively. So the mean value of $|x_j|^2$ is

$$E\left(|X_j|^2\right) = E\left(|S_j + N_j|^2\right) = E\left(|S_j|^2\right) + E\left(|N_j|^2\right) + 2E\left(\mathrm{Re}\{S_j N_j^*\}\right) \quad (13)$$

In PMC method, it is assumed that the speech signal and the noise are uncorrelated, therefore the correlation term $E\left(\mathrm{Re}\{S_j N_j^*\}\right)$ in eq (13) equals zero. This is the way to arrive at a tractable solution. However, this assumption may not necessarily be true in short-term analysis for feature extraction, in which the noise is certainly not stationary, and usually not zero mean. In such cases, the nonzero correlation term $E\left(\mathrm{Re}\{S_j N_j^*\}\right)$ may not be negligible, especially under low SNR conditions.

Due to lack of the complete information in estimating the correlation term $E\left(\mathrm{Re}\{S_j N_j^*\}\right)$ in PMC processes, the estimation may be performed by approximation as follows. First,

$$\mathrm{Re}\{S_j N_j^*\} = |S_j||N_j|\cos\left(\theta_{S_j} - \theta_{N_j}\right) \quad (14)$$

where $S_j = |S_j|\exp\left(j\theta_{S_j}\right)$ and $N_j = |N_j|\exp\left(j\theta_{N_j}\right)$

so

$$E\left(\mathrm{Re}\{S_j N_j^*\}\right) = E\left(|S_j||N_j|\cos\left(\theta_{S_j} - \theta_{N_j}\right)\right) \quad (15)$$

If we assume that $|S_j|$, $|N_j|$ and $\cos\left(\theta_{S_j} - \theta_{N_j}\right)$ are mutually independent, then eq (15) becomes

$$E\left(\mathrm{Re}\{S_j N_j^*\}\right) = E\left(|S_j|\right)E\left(|N_j|\right)E\left(\cos\left(\theta_{S_j} - \theta_{N_j}\right)\right) \quad (16)$$

Furthermore, because of the inequality

$$E\left(Y^2\right) \geq [E(Y)]^2,$$

we can have

$$E\left(|S_j|\right) = \alpha_j\left[E\left(|S_j|^2\right)\right]^{\frac{1}{2}}$$

and

$$E\left(|N_j|\right) = \beta_j\left[E\left(|N_j|^2\right)\right]^{\frac{1}{2}}$$

where $0 \leq \alpha_j, \beta_j \leq 1$, then eq (16) can be rewritten as then

$$E\left(\mathrm{Re}\{S_j N_j^*\}\right) = \alpha_j\beta_j E\left(\cos\left(\theta_{S_j} - \theta_{N_j}\right)\right)\left[E\left(|S_j|^2\right)E\left(|N_j|^2\right)\right]^{\frac{1}{2}}$$

$$= \gamma_j\left[E\left(|S_j|^2\right)E\left(|N_j|^2\right)\right]^{\frac{1}{2}} \quad (17)$$

where γ_j is defined as

$$\gamma_j = \alpha_j\beta_j E\left(\cos\left(\theta_{S_j} - \theta_{N_j}\right)\right). \quad (18)$$

Because the terms $E\left(|S_j|^2\right)$ and $E\left(|N_j|^2\right)$ are both known parameters, the only variable to be determined is γ_j. In the next section, we will show that by proper choice of γ_j, the performance of the PMC compensated HMM can be improved. Hence, the part different from the original PMC method is in the composition process, i.e. in eq (5),

$$\hat{\mu}_j = \mu_j + g\tilde{\mu}_j + 2\gamma_j\sqrt{\mu_j \bullet g\tilde{\mu}_j} \quad (19)$$

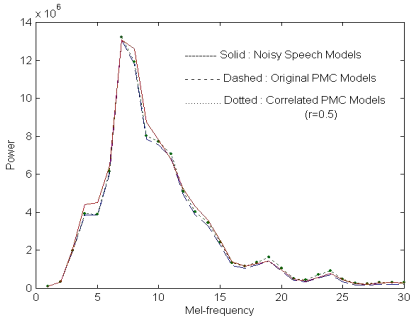
for each component of $\hat{\boldsymbol{\mu}}$.

4. Experimental Results

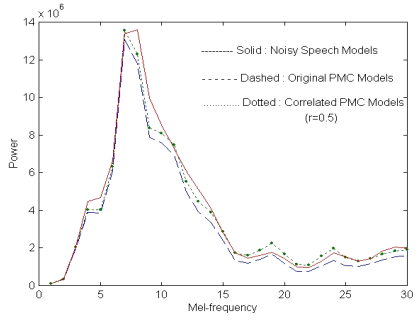
Some preliminary experiments were performed to verify the concept mentioned here. The training speech database used in the experiments contains 3 sets of 1345

isolated syllables in Mandarin Chinese produced by a speaker. It is used to train 113 right context-dependent (RCD) INITIAL HMM's and 41 context-independent (CI) FINAL HMM's. Another set of 1345 syllables produced by the same speaker is used as the test data to be recognized in speaker dependent mode. 14 order mel-frequency cepstral coefficients are used as the feature parameters. Also, the continuous density HMM (CHMM) is trained with 1 state per INITIAL model, 2 states per FINAL model and 2 mixtures per state. Noise HMM's for different levels of white noise to be added into the clean speech are also individually trained, composed of one state and one mixture per state. Furthermore, HMM's are also trained with the noise corrupted speech data.

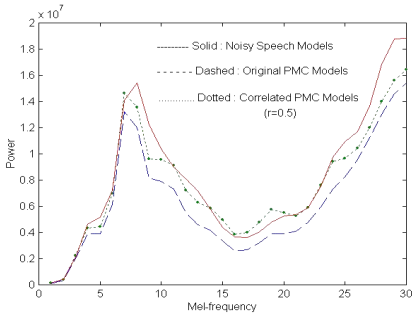
We first examine the mismatch of the original PMC-based HMM's and the correlated PMC HMM's proposed here with respect to HMM's trained with noisy speech. Figure 1 shows the spectral envelopes for the 3 sets of models averaged over all the 113 INITIAL models and 41 FINAL models at SNR of 30, 20 and 10 dB. The mean and standard deviation values of the differences between the spectral envelopes for the two versions of PMC-based HMM's and the HMM's trained with noisy speech are also listed in Table 1. Apparently, the deviations between both PMC-based models and the models trained with noisy speech are different at different mel-frequency values. From both the figure and the table, it's obvious that the difference between the models increases as the SNR becomes worse. Since the term $E\left(\mathrm{Re}\{S_j N_j^*\}\right)$ increases with $E(|N_j|^2)$ in Eq (17), it is reasonable to say that the deviations may be due to, at least partially, the lack of good estimation of the term $E\left(\mathrm{Re}\{S_j N_j^*\}\right)$. It is also clear from both the figure and the table that the correlated PMC model proposed here made reasonable estimates of the term $E\left(\mathrm{Re}\{S_j N_j^*\}\right)$, thus resulted in relatively smaller deviations as compared to the original PMC.



(a) SNR=30dB



(b) SNR=20dB



(c) SNR=10dB

Figure 1. The spectral envelopes under various SNR

	Original PMC HMM's		Correlated PMC HMM's	
SNR	mean	std.	mean	Std.
30dB	$2.00\text{x}10^5$	$2.50\text{x}10^5$	$7.82\text{x}10^4$	$2.51\text{x}10^5$
20dB	$5.29\text{x}10^5$	$4.82\text{x}10^5$	$1.44\text{x}10^5$	$4.58\text{x}10^5$
10dB	$1.79\text{x}10^6$	$1.17\text{x}10^6$	$5.73\text{x}10^5$	$1.11\text{x}10^6$

Table 1. The mean and standard deviation values of the differences between the spectral envelopes for the two versions of PMC-based HMM's and the HMM's trained with noisy speech

The next test is on the model distance measures using the averaged Kullback-Leibler (KL) number. For Gaussian distributed variables, the KL number is defined as [5]:

$$D_{KL}(p,q)=\frac{1}{2}\left[\log\left(\frac{\Sigma_q}{\Sigma_p}\right)+\frac{\left(\mu_p-\mu_q\right)}{\Sigma_q}+\left(\frac{\Sigma_p}{\Sigma_q}-1\right)\right]$$

where p is the true distribution and q is the estimated distribution. The model distance is then obtained by averaging all the KL numbers between each pair of corresponding Gaussian distributions on a per feature vector component basis. For the correlated PMC method proposed here, the correlation parameter γ_j was chosen as 0.5. The average KL numbers for clean speech HMM's, original PMC HMM's and the correlated HMM's proposed here with respect to the noisy speech HMM's at different SNR values are shown in Table 2 and Fig. 2. It is clear from Table 2 that the original PMC method can significantly reduce the model distance, while in all cases the correlated PMC method proposed here can further

reduce the model distance to a certain extent. This is also clear from Fig. 2.

SNR[dB]	30	25	20	15	10
Clean	0.513	0.771	1.087	1.429	1.819
PMC HMM	0.107	0.183	0.160	0.154	0.165
Correlated PMC γ_j =0.5	0.099	0.160	0.149	0.145	0.158

Table 2. Averaged KL numbers between the different versions of models and the noisy speech model.

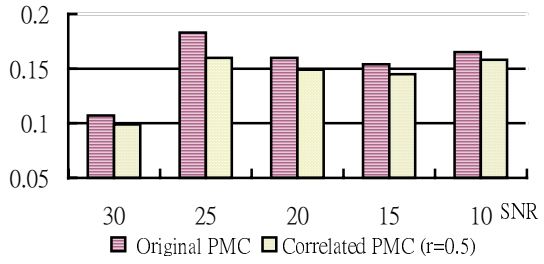


Figure 2. The comparison of averaged KL numbers for original PMC and the correlated PMC proposed here.

SNR (dB)	Clean speech HMM's	Noisy Speech HMM's	Original PMC HMM's	Correlated PMC HMM's			
				γ_j =0.3	γ_j =0.5	γ_j =0.7	γ_j =0.9
30	48.33	80.67	74.28	76.88	77.55	77.25	78.36
25	33.09	77.32	63.57	69.29	69.52	69.44	69.44
20	14.42	71.15	49.89	56.73	57.55	57.03	59.26
15	5.13	60.74	36.80	45.80	46.69	46.62	46.77
10	2.01	49.74	27.66	36.65	36.88	37.25	36.80

Table 3. Recognition accuracies using different versions of models

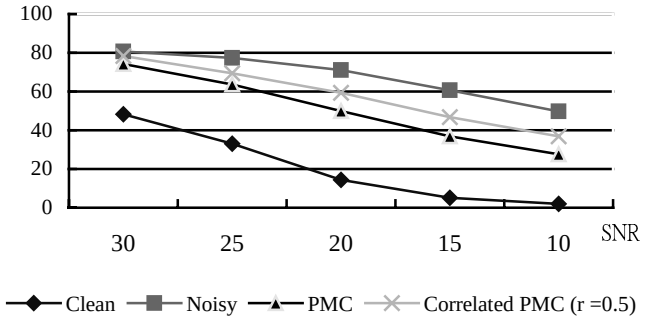


Fig.3 Recognition accuracies using different versions of models

Finally, the recognition accuracies using different versions of models are compared in Table 3 and Figure 3. One can find in the second column that the recognition accuracies are seriously degraded using the clean speech HMM's, especially when SNR becomes worse. However, when the noisy speech HMM's are used, the recognition rates (in the third column of Table 3) can be significantly improved, but it is time-consuming and cost-intensive in training these noisy speech models. In the original PMC models, only a short period of noise is used, thus is practically much feasible. As can be found in the fourth column of Table 3, the recognition performance is improved significantly when SNR is high, but the improvements are reduced when SNR becomes lower. However, the correlated PMC proposed here (listed in the last four columns) always provides better compensation compared to the original PMC, especially in the low SNR cases. The accuracies now are actually much more closer to the results of the noisy speech models in the third column. For instance, in comparison with the original PMC, the choice of γ_j =0.5 in correlated PMC reduces the error rate by 12.71%, 16.33%, 15.29%, 15.65% and 12.75% for SNR values of 30dB, 25dB, 20dB, 15,dB and 10dB respectively. Clearly the added nonzero correlation term in eq (17) not only reduces the model distances but also improves the recognition accuracy. As was shown in the experimental results, by simple but proper assignment of the value γ_j, the performance of the compensated HMM can be significantly enhanced. It is believed that the recognition performance can be further improved with more careful choice of the parameter value.

5. Conclusion

In this paper, it is shown that better estimation of the correlation term in the original PMC method can produce much better acoustic models and accuracies for speech recognition under noisy conditions, especially when SNR is low. It is believed that the performance of the PMC method can be further improved if more precise estimation of the correlation can be applied.

Reference:

[1] M.J. Gales and S. J. Young, "Robust Speech Recognition in Additive and Convolutional Noise Using Parallel Model Combination", Computer, Speech and Language 9, pp. 289-307,1995.

[2] M.J. Gales and S. J. Young, "An Improved Approach to the Hidden Markov Model Decomposition of Speech and Noise", IEEE, ICASSP'92, I, pp. 233-236, 1992.

[3] H. Yamamoto, M. Yamada and T. Kosaka, "Independent Calculation of Power Parameters on PMC Method", IEEE, ICASSP'96, pp. 41-44.

[4] N. B. Yoma, F. McInnes and M. Jack, "Weighted Matching Algorithms and Reliability in Noise Cancelling by Spectral Subtraction", ICASSP'97, pp. 1171-1174.

[5] B. H Juang, and L. R. Rabiner, , "A Probabilistic Distance Measure for Hidden Markov Models", AT&T Technical Journal 64, pp. 391-408.