A Novel Use of Residual Noise Model for Modified PMC

Cailian Miao Yangsheng Wang

National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
{clmiao,wys}@mail.pattek.com.cn

ABSTRACT

In this paper, a new approach based on model adaptation is proposed for acoustic mismatch problem. A specific bias model—residual noise model—is presented, which is the joint compensation model for additive and convolutive bias. The novel noise model is estimated on the basis of maximum likelihood manner. In conjunction with the Parallel Model combination (PMC), it is effective for noisy environments. The experiments have been done based on the Cambridge's HTK toolkit to implement the continuous Mandarin digit recognition in noisy environments.

1. Introduction

The robustness of speech recognition techniques has been an unbroken interest in this field for decades. The robustness techniques can be categorized into 3 classes: using robust features and/or distortion measures so that it is less sensitive to the various acoustic conditions, the MFCC, SMV, OSALPC coefficients are all typical examples in this class; applying enhancement methods to recover the noisy observations to clean speech as possible, the spectral subtraction, FCDCN, RATZ are typical examples; adapting the model to describe the noisy signal better, the noise masking, SND, VTS, MLLR, PMC are typical examples. The PMC method has been one of the most popularly used approaches in the last class.

The mismatch between training and testing can be attributed to channel distortion and ambient noise, they can be regarded as convolutive bias and additive bias in linear spectral domain. PMC has been used in many applications to handle both types of noise successfully. However, there are still some limitations: First, even with precise noise models, it is believed that some approximation and assumptions made in domain transformation and parameter combination processes are not accurate enough, which may degrade the performance of PMC. Second, the lower the SNR of noisy observation, the more dominant the additive noise model, the combined models should be noise models rather than noisy speech models, it is hard to get satisfactory result with such additive noise models.

In second robustness class, some efficient signal enhancement techniques, such as spectral subtraction (SS), cepstral mean normalization (CMN), have been widely used to reduce the effect of noise. All of them assume that the residual noise is small enough so that its effect can be ignored. However, the deviations of the distributions of the enhanced speech from those of the corresponding clean speech are observed. This fact implies that residual noise requires to be modeled. The useful approaches combine such specific model with clean speech models to match the enhanced test data better.

Motivated by these reasons, the new PMC is proposed. First, the enhancement method is used to restore the noisy speech signal to clean speech as close as possible; then the model adaptation method is applied so that the combined models can match the enhanced speech better. New residual noise model and pseudo clean speech models are introduced into PMC that can incorporate the signal enhancement with environment adaptation strategies to increase the robustness and improve the performance in noisy conditions.

In our experiment, Cambridge's HTK toolkit 3.0 was used as test platform with suitable modification embedding the new PMC algorithms to implement the continuous Mandarin digit recognition. The training data were collected in clean office environment while the test data included the data contaminated by white Gaussian noise at different SNR levels and different types of convolutive noise. The results of experiment clarify the effectiveness of the proposed approach.

The paper is organized as following: In Section 2, the fundamental of PMC is briefly reviewed. In Section 3, the processing of the joint bias compensation is introduced and the ML bias estimation in an expectation maximization framework is formulated. In Section 3, the experimental results are presented. Finally, the future work is suggested and the conclusion is drawn.

2. The FUNDAMENTAL of PMC

The PMC approach is briefly summarized here to help the discussion below.

It defines a proper mismatch function to represent the effect that the noise has on the elements of the speech feature vector. In many situations, both additive and convolutive noise is present. For static features:

$$y^l(\tau) = \log\left(\exp\left(x^l(\tau) + h^l(\tau)\right) + \exp\left(n^l(\tau)\right)\right) \quad (1)$$

where $x(\tau)$、$n(\tau)$、$h(\tau)$、$y(\tau)$ is the clean speech feature, additive noise feature, convolutive noise feature and noisy speech feature in τ frame, respectively. The superscript l is a parameter in log spectral domain. Similarly superscript c is the parameters in cepstral domain. All models in this paper are set mixture Gaussian distributions, μ is the mean vector，Σ is the covariance vector, i is the dimension index.

* The Conference Participation is supported by Nokia Bridging the World Program

In PMC approach, the models should be transformed to the log spectral domain by invert DCT transformation, then to linear domain by powering the component, the additive noise and speech is combined. At last, the inverse transformation can be used to restore the models to the cepstral domain.

Some approximation and assumptions are made in domain transformation and parameter combination processes above. For example, in the linear domain the distributions are assumed to be log-normal. Meantime, the sum of two log-normal variables is approximated to be also log-normal distribution, which may degrade the performance of PMC in some actual situations.

3. Joint Compensation Algorithm

Since it is hard to find a simple mismatch function to describe the enhanced speech, the PMC approach has never incorporated with enhancement schemes yet. The innovative attempt has done in this paper.

The procedure is: First, the enhancement method is used to restore the noisy speech signal to clean speech as possible; then, the model adaptation method is applied so that the combined models can match the enhanced speech better.

With the modified PMC, the residual noise is the joint bias compensation for additive and convolutive noise; the bias parameter combinations are performed only one time in cepstral domain without domain transformation, which is based on the new mismatch function. Thus the limitations of traditional PMC are overcome.

The key point of the new PMC is to find effective enhancement approaches and define proper mismatch function, also to model the residual noise precisely and combine the parameters.

3.1 Finding Enhancement Approaches

There are many enhancement approaches which proved to be effective. In this paper, the modified Spectral Subtraction is used.

SS scheme can modify the magnitude of signal in power spectra by using a filter as following:

$$\left|\hat{X}(\omega)\right| = \left|Y(\omega)\right| \cdot H_{ss}(\omega) \tag{2}$$

where $\left|\hat{X}(\omega)\right|$ is the enhanced observation data. ω is frequency bin, $H_{ss}(\omega) = \sqrt{\max(1 - 1/SNR(\omega), a)}$ is an SNR-based gain function to suppress frequency components with low SNR. a is the threshold value. $SNR(\omega) = \left|Y(\omega)\right| \Big/ \left|\hat{N}(\omega)\right|$, $\hat{N}(\omega)$ is the estimation of the noise spectrum that is obtained by picking the minimum value from consecutive values of the estimation of noisy speech power spectrum $\hat{P}(\omega,\tau) = \beta\hat{P}(\omega,\tau-1) + (1-\beta)\left|Y(\omega,\tau)\right|^2$, where β is the weight coefficient, τ is the frame index.

With such approach, the output data has significantly less noise, though it exhibits that called musical noise, which is caused by discontinuity of the $H_{ss}(\omega)$ function. The residual noise contains not only the remained additive noise, but also the new musical noise.

3.1. Defining Mismatch Function and combining the parameter

It is known that the residual noise remains even when the enhancement approaches are applied to the noisy speech, which degrades the performance. In this paper, the mismatch function is defined as following:

$$\hat{x}^l(\tau) = \log\left(\exp\left(x^l(\tau) + r^l(\tau)\right)\right) \tag{3}$$

where $\hat{x}(\tau)$, $r(\tau)$ is the enhanced (pseudo clean) speech and the residual noise parameter, respectively.

The effect between the residual noise and the clean speech signal is approximated as linear addition in cepstral domain. The parameter combinations are shown as following:

$$\mu_{\hat{x}}^c = \mu_x^c + \mu_r^c \tag{4}$$

$$\Sigma_{\hat{x}}^c = \Sigma_x^c + \Sigma_r^c \tag{5}$$

With such residual noise model, the joint bias for additive and convolutive noise is performed. The bias value of residual noise can be estimated with maximum likelihood algorithm.

3.2. Modeling the Residual Noise

Given the enhanced feature set $\hat{X}$,the basic assumption is at the state level, the $\hat{X}$ is related to the clean speech feature X by the equation:

$$\hat{X} = X + B \tag{6}$$

where B is the feature of a residual noise model . Furthermore, X and B is independent.

It is required to determine the residual noise model parameters $\lambda_b \equiv (\mu_b, \Sigma_b)$ .Given the enhanced feature set $\hat{X}$ and clean speech model λ_X , the solution using maximum likelihood is shown as following:

$$\lambda_b' = \arg\max_{\lambda_b} P\left(\hat{X}\middle|\lambda_x, \lambda_b\right) \tag{7}$$

The above ML parameter estimation can be solved by EM algorithm. In the first E-step, the following auxiliary function is calculated as shown below:

$$\begin{aligned} Q(\lambda_b'|\lambda_b) &= E\left[\log p\left(\hat{X}, S, K\middle|\lambda_b', \lambda_x\right)\middle|\hat{X}, \lambda_b, \lambda_x\right] \\ &= \sum_{S,K} P(S,K|\hat{X}, \lambda_b, \lambda_x) E\left[\log P(X, B, S, K|\lambda_b', \lambda_x)\middle|\hat{X}, S, K, \lambda_b, \lambda_x\right] \end{aligned} \tag{8}$$

where S , K is state and mixture component sequences of an HMM. Ignoring some irrelevant terms, we can get:

$$\begin{aligned} Q(\lambda_b'|\lambda_b) &= -1/2 \sum_t \sum_i \sum_j \gamma_t(i,j) \\ &\cdot E\left[\left(\log\left|\Sigma_{b,i}\right| + (b_t - \mu_{b,i})^T \Sigma_{b,i}^{-1}(b)(b_t - \mu_{b,i})\right)\middle| y_t, i, j, \lambda_b, \lambda_x\right] \end{aligned} \tag{9}$$

where $\gamma_t(i,j) = P(s_t = i, k_t = j | y_t, i, j, \lambda_b, \lambda_x)$ is a posteriori probability of the occupying state i and mixture component j at time t .

In the second M-step, the auxiliary function is differentiated with respect to μ_{bi} and Σ_{bi} , and equating to zero, we get:

$$\mu_{bi} = \sum_t \sum_j \gamma_t(i,j) E\left[b_t\middle|y_t, i, j\right] \Big/ \sum_t \sum_j \gamma_t(i,j) \tag{10}$$

$$\Sigma_{bi} = \sum_t \sum_j \gamma_t(i,j) E\left[b_t b_t^T\middle|y_t, i, j\right] \Big/ \sum_t \sum_j \gamma_t(i,j) - \mu_{bi}\mu_{bi}^T \tag{11}$$

Thus applying the EM step iteratively, the bias estimation can be gotten from some initial bias parameter values. Each iteration ensures the increase of occupying probability of the enhanced speech likelihood.

The whole procedure is shown in Fig.1

However, Strictly speaking, it is difficult to find closed-form solution for the expected value $E\left[b_t \middle| y_t, j, i\right]$ and $E\left[b_t b_t^T \middle| y_t, j, i\right]$. In this paper, the covariance values are assumed to be diagonal, so the initial bias is approximated as following:

$$E\left[b_t \middle| \hat{x}_t, i, j\right] \approx \hat{x}_t - \mu_{x,i,j} \tag{12}$$

$$E\left[b_t^2 \middle| \hat{x}_t, i, j\right] \approx \left(\hat{x}_t - \mu_{x,i,j}\right)^2 + \sigma_{x,i,j}^2 \tag{13}$$

The iterative expected values can be equated as following:

$$E\left[b_t \middle| \hat{x}_t, i, j\right] \approx \rho_{i,j}\mu_{b,i} + \left(1 - \rho_{i,j}\right)\left(\hat{x}_t - \mu_{x,i,j}\right) \tag{14}$$

$$E\left[b_t^2 \middle| \hat{x}_t, i, j\right] = E^2\left[b_t \middle| \hat{x}_t, i, j\right] + \sigma_{x,i,j}^2\sigma_{b,i}^2 \Big/ \left(\sigma_{x,i,j}^2 + \sigma_{b,i}^2\right) \tag{15}$$

The experiment has shown the effectiveness of modified PMC below.

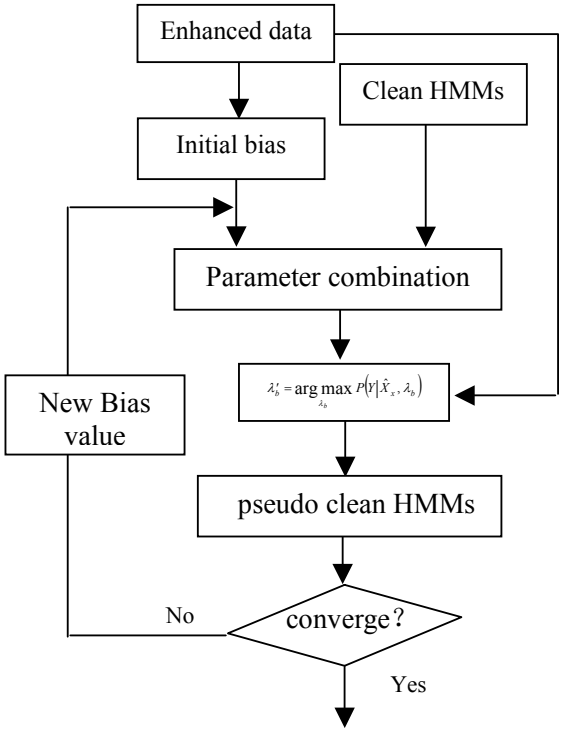


Figure 1. The diagram of ML estimation for residual noise model.

In such way, the joint compensation algorithm for additive and convolutive bias is proposed, and the PMC can be in conjunction with enhancement schemes, which will improve the robustness of system. The ML is used to estimate the specific residual noise model.

4. Experiments

The experiments have been performed to assess the proposed approaches. The clean speech database consists of 3200 sentences uttered by 40 speakers (80 sentence per person) to implement the speaker independent continuous mandarin digit numbers (containing 1~8 numbers per sentence) recognition. Models are trained for one group and tested on the other group. We tried three different combinations of the group and the average results are reported. Each train set consists of 80*36 trails.

The pre-emphasized feature vector of the speech data is described by the MFCC including 26 vectors, (13 MFCC, including zero coefficient, their delta components obtained by linear regression filter) windowed in 25 ms Hamming in 10 ms step with 26 filter banks. Each phone model has 7 states with single mixture Gaussian distribution. The accuracy rate is above 91 in clean environment.

In our experiments, the test sentences are the clean speech corrupted by additive noise in various levels and (or) types of convolutive noise. The additive noise is white Gaussian noise. Each convolutive noise contains different sample lengths: 50, 100, 150, 200, 400,which is corresponding to the frequency response time of transducers or communication channels.

Table.1. Recognition results (Word accuracy rate %) for the speech corrupted by additive and (or) convolutive noise, applying the spectral subtraction scheme and the PMC approaches.

Conditions		baseline	SS	PMC
additive noise	SNR=12dB	34.08	35.00	43.18
	SNR=8dB	12.64	16.50	35.54
	SNR=4dB	8.40	8.40	15.81
	SNR=2dB	8.40	8.40	8.40
convolutive noise (samples)	50	34.85	29.85	67.15
	100	32.69	31.92	66.15
	20	27.83	28.45	60.83
	400	21.28	21.43	38.40
additive + convolutive	SNR=10dB	16.50	16.59	20.66
	SNR=5dB	13.72	14.03	20.97
	SNR=2dB	9.95	9.99	16.58

The results of the spectral subtraction approach, traditional PMC for different noisy conditions are compiled in Table.1 and the baseline is also shown for comparison. The additive noise ranges from 2~12dB, the samples' lengths of convolutive noise are 50,100,200,400. Both additive and different patterns of convolutive noises are added to speech in the last 3 rows that degrade SNR below 10 dB. Where the results of "baseline" show the noise effect severely degrades the recognition performance. The "SS" shows the results of using spectral subtraction approach. The "PMC" indicates the results of applying the conventional PMC, which the additive or (and) convolutive noise models are trained preliminarily.

Table1 shows the PMC and SS are all effective to increase the robustness of system in most cases, even in PMC the noise models are assumed known. Although in low SNR (SNR=2dB), PMC can not improve the recognition rate, which shows the limitation of PMC. On the other hand, Spectral Subtraction scheme are not effective in convolutive noise.

The same tests are performed to evaluate the modified PMC. 20 sentences cleaned by Spectral Subtraction are used as the adaptation data. At first, a global mean bias value of residual noise model is estimated using the statistical scheme as following:

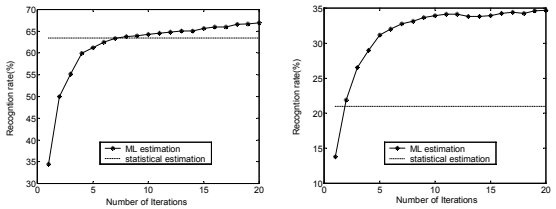
$$\mu_R^{cep} = \arg\min_{b_R^{cep}} \left[\left| \sum_x \sum \mu_x^{cep} - 1/T \sum_{\tau=1}^{T} \hat{x}^{cep}(\tau) \right| \right] \tag{16}$$

The result of modified PMC is shown in Table.2.

Table.2. Recognition results for speech corrupted by additive and (or) convolutive noise applying the new PMC approaches.

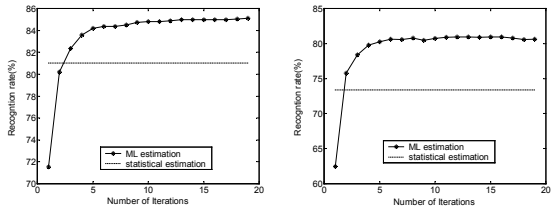
Conditions		PMC	SS+PMC
additive noise	SNR=12dB	43.18	64.30
	SNR=8dB	35.54	54.51
	SNR=4dB	15.81	35.47
	SNR=2dB	8.40	20.97
convolute noise (samples)	50	67.54	81.65
	100	66.15	81.03
	200	60.13	77.26
	400	38.40	73.40
additive + convolute	SNR=10dB	20.97	62.76
	SNR=5dB	20.66	52.43
	SNR=2dB	16.58	39.24

The "PMC+SS" presents the result when we apply SS as preprocessing procedure for PMC and the residual noise model used instead of additive and convolutive noise model. The accuracy rate of "PMC+SS" is significantly more than those of PMC and SS, especially in low SNR or adverse conditions. For example, in case of SNR=12dB the modified PMC achieves more than 21 higher accuracy. At least in our testing conditions, the modified PMC using residual noise model can lead to high improvement in performance in adverse environments.



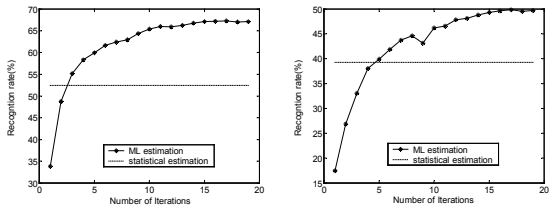
(a) SNR=12dB (b) SNR=2dB

Figure 2. The recognition results of new PMC (Word accuracy rate %) in additive noise conditions, estimating the residual noise model on the iterative ML manner.



(a) 100 samples (b) 400 samples

Figure 3. The recognition results of new PMC in convolutive noise conditions, estimating the residual noise model on the iterative ML manner.



(a) SNR=10dB (b) SNR=2dB

Figure 4. The recognition results of new PMC in additive & convolutive noise conditions, estimating the residual noise model on the iterative ML manner.

In order to improve the performance further, the bias parameters is estimated using Eq.(10)~Eq.(15). Fig.2 shows the results of the modified PMC using such bias in the three mismatched conditions above with corresponding to iterative times. The recognition results of previous modified PMC using Eq.(16) are also shown for comparison.

Where the solid lines with "*" shows the result of modified PMC using the bias parameters estimated with ML, the dash-dot line shows the result of modified PMC using statistical estimation of bias with Eq.(16).

All the presented results are significantly better than those with statistical bias parameters when some iteration have been done. For example, In SNR=2dB additive noise condition, the recognition rate with global static bias compensation is only 20.97%. Application of ML estimation for bias compensation can improve the result to 34.62% when performing 19 iterations. In convolutive condition (400 samples), the former is 73.40 % while the latter is 80.65% for 19 iterations. In adverse conditions (both additive and convolutive noise existing), when SNR=5dB, the former is 52.43%, the latter is 67.08%; when SNR=2dB, the former is 39.24% while the latter can achieve to 49.73%. The results illustrate that the performance of the combined models, which are composed of iteratively estimated bias parameter, is highly improved in adverse situations.

The results above show the effectiveness of the modified PMC.

5. Conclusions

With some speech enhancement schemes, there is still some residual noise remained which degrades the performance more or less. In this paper, the residual noise is modeled and used in PMC approach to increase the robustness. With such new approach, some limitations of PMC are overcome in some extent. From the experiment with different convolutive patterns and additive conditions, the proposed PMC algorithm significantly outperforms both speech enhancements and traditional PMC. In a word, the PMC proposed in this paper provides an effective approach to make speech recognition more robust.

6. References

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

[2] D.Mansour and B.H.Juang, "The short-time modified coherence representation and noisy speech recognition", IEEE trans. On signal processing, Vol.37, June, 1989

[3] J.Hernando and C.Nadeu, "speech recognition in noisy car environment absed on OSALPC representation and robust similarity measuring techniques", Proc –ICASSP '2001,p69~72, 2001

[4] Harald Gustafsson, Sven Erik Nordholm, Ingvar Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging", IEEE trans. On speech and audio processing, Vol.9, No.8, November, 2001