A Comparative Analysis of Blind Channel Equalization Methods for Telephone Speech Recognition

Wei-Wen Hung and Hsiao-Chuan Wang
Department of Electrical Engineering, National Tsing Hua University,
Hsinchu, Taiwan, 30043, Republic of China
E-mail : hcwang@ee.nthu.edu.tw

Abstract

A blind channel equalization method called signal bias removal (**SBR**) has been proposed and proved to be effective in compensating the channel effect in telephone speech recognition. However, we found that the SBR method didn't work well when additive noise and multiplicative distortion are taken into account at the same time. In this paper, we propose a new method called modified signal bias removal (**MSBR**) which tries to overcome the problem described above in the SBR method. Some experiments are conducted to evaluate the effectiveness of the MSBR method. Experimental results show that the MSBR method outperforms the SBR no matter additive noise is considered or not in a telephone speech recognition system.

1. Introduction

When a speech recognition system based on HMM is deployed over a telephone network, undesired effects due to adverse interferences will make noisy speech signals and reference models mismatched and cause serious degradation in recognition performance. To realize a robust speech recognizer, the problem of how to restore original speech signals from the contaminated speech transmitted over a telephone network has to be solved. Mokbel et al. [1] proposed a cepstral mean subtraction (CMS) method to reduce the channel effect of a telephone line. Hermansky et al. [2] used the RelAtive SpecTrAl (RASTA) for the compensation of steady-state spectral distortions in a telephone channel. Codeword-dependent cepstral normalization (CDCN) and SNR-dependent cepstral normalization (SDCN) were also proposed for reducing the variability of additive noise and channel effect [3]. Besides, the MAP channel estimation methods [4] which based on a prior channel statistics were successfully applied for telephone speech recognition. From a practical point of view, many of these compensation algorithms which take advantage of the availability of a priori knowledge about the testing environment are less attractive than those that require no any previous information, i.e., the blind equalization methods [5]-[8].

A promising blind channel equalization method called signal bias removal (SBR) is proposed by Rahim and Juang [5]-[7]. In this method, undesirable components due to unknown effects in a telephone system are minimized by using the maximum likelihood estimation. The results of their experiments in which only a multiplicative spectral bias was considered demonstrated the effectiveness of the proposed method. However, we found that this blind channel equalization method can not work well when the speech signal mixed with additive noise is transmitted over a telephone channel. In this paper, a new approach based on local maximum likelihood estimation is proposed to improve the efficiency of minimizing those undesirable effects in telephone speech recognition.

2. Effectiveness of SBR Method

The SBR method reported in the paper [5]-[7] only considers an additive bias, i.e., a multiplicative spectral distortion. When an additive noise is also taken into account, we found that the efficiency of the SBR method drops.

The formulation of the SBR method is to maximize the likelihood function which is based on the set of all word models, and can be defined as

$$p(X|\Lambda) = \max_{\Lambda(w)} p(X|\Lambda(w)). \qquad (1)$$

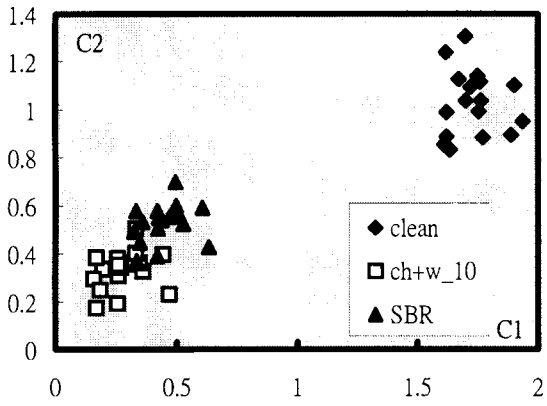
In above equation, the likelihood function based on the hidden Markov model of word w is formulated as

$$p(X|\Lambda(w)) = \prod_t \max_i p(x_t|\Lambda_i(w)), \qquad (2)$$

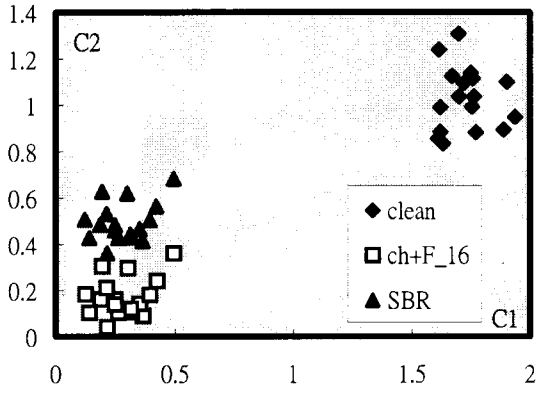
where $X=\{x_1,x_2,\cdots,x_t,\cdots,x_T\}$ is an observation sequence with T frames. $\Lambda = \{\Lambda(w),\ w = 1,2,...,N\}$ represents the set of all word models, and $\Lambda(w) = \{\Lambda_i(w),\ i = 1,2,...,M_w\}$, where N is the number of word models, M_w is the number of states in the w-th word model $\Lambda(w)$, and $\Lambda_i(w) = N(\mu_i(w), \sum_i(w))$ is a Gaussian probability density function (pdf). The contaminated signal $Y=\{y_1,y_2,\cdots,y_t,\cdots,y_T\}$ can be modeled by adding a bias vector as

$$y_t = x_t + b \qquad t = 1,2,\cdots,T. \qquad (3)$$

Based on the set of all word models Λ, we can



(a) for white noise with SNR at 10 dB.



(b) for F16 colored noise with SNR at 10 dB.

Fig. 1 The movement of cepstral clusters under various conditions.

determine the maximum likelihood bias estimator $\bar{b}$ by using the following equation

$$\bar{b} = \frac{1}{T}\sum_{t=1}^{T}(y_t - \delta_t), \tag{4}$$

where $\delta_t = \arg\max_{\Lambda_i(w^*)} p(y_t - b|\Lambda_i(w^*)),$ (5)

and $w^* = \arg\max_{w} p(Y|b,\Lambda(w)).$ (6)

From above description, it is evident that the maximum likelihood estimator of the bias signal $\bar{b}$ is based on the set of all word models, i.e., a global maximization range.

Here is an experiment to show the effectiveness of SBR method. An isolated utterance of Mandarin digit '5' is corrupted by a white Gaussian noise and a F16 colored noise with signal-to-noise ratio (SNR) at 10 dB. Then the noise contaminated utterance is passed through a simulated telephone channel. Every feature vector is represented by a 12-order cepstral coefficients derived from linear predictive (LP) analysis. Those feature vectors generated at the output of a channel filter are compensated by using the SBR method in which the bias removal procedure is repeated five times in the cepstral domain. By projecting the feature vectors onto a C1-C2 plane formed by the first two cepstral coefficients, Fig.1 shows the scattering plots of the cepstral coefficients we generated under different experimental conditions. From Fig.1, we can observe that the SBR method can reduce the distances between the cepstral clusters (labeled by 'ch+w_10' and 'ch+F_16') and the clean cepstral cluster.

3. Modified Signal Bias Removal Method

Vaseghi and Milner [8] had proposed a hypothesized maximum likelihood algorithm (HML) to improve the SBR method. The procedure of HML method is summaried as follows :

For $w = 1$ *to* N (number of word models)

{ *step 1.* Basing on ML criterion and using HMM, $\Lambda(w)$, to estimate the channel, $\hat{h}_w$,

step 2. Using $\hat{h}_w$ to estimate the input as $\hat{X}(\hat{h}_w) = Y - \hat{h}_w$,

step 3. Computing a score for model $\Lambda(w)$, given the estimated input $\hat{X}(\hat{h}_w)$.

}

step 4. Selecting the most probable word.

Here we propose an alternate method called modified signal bias removal method (MSBR). The idea is that the contaminated signal can be evaluated independently on every word model to estimate a mean bias corresponding to different HMM. The contaminated signal should be first compensated by this mean bias and then use the result to compute a likelihood score based on the corresponding word model. The mean bias which has the highest likelihood score is considered as the most reliable and probable one.

There are some differences between MSBR and HML In the MSBR method, only one most probable bias estimator is used to calculate the likelihood scores for the whole word models. However, in the HML method, the likelihood score with respect to a word model is based on different bias estimator for every word model. The detailed formulation of our approach is shown in Fig. 2. First, the contaminated signal Y is evaluated independently on every word model in order to obtain a mean bias $\bar{b}_k$ corresponding to word model $\Lambda(k)$, i.e.,

$$\bar{b}_k = \frac{1}{T}\sum_{t=1}^{T}(y_t - \delta_{k,t}) \quad k = 1,2,\cdots,N, \tag{7}$$

where $\delta_{k,t} = \arg\max_{\Lambda_i(k)} p(y_t - b_k|\Lambda_i(k)).$ (8)

Once the mean bias $\bar{b}_k$ is obtained, the contaminated signal is compensated by subtracting $\bar{b}_k$ from the signal

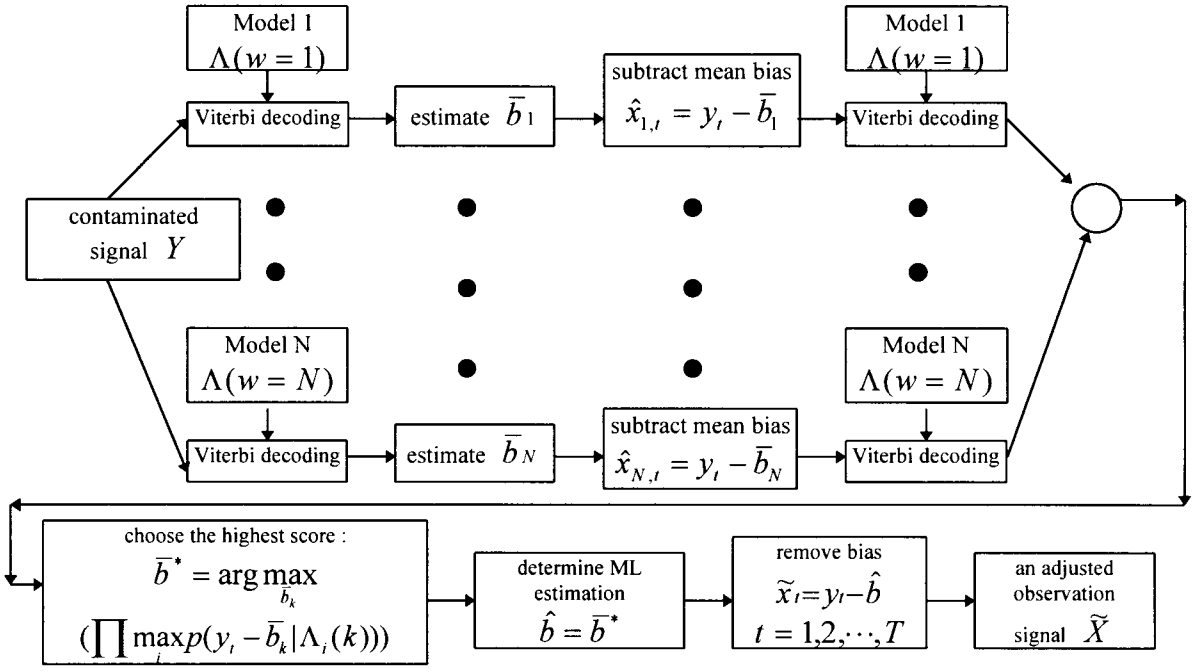


Fig.2 modified signal bias removal (MSBR) method

y_t to get a bias-removed signal $\hat{x}_{k,t}$,

$$\hat{x}_{k,t} = y_t - \bar{b}_k \quad k=1,2,\cdots,N, \quad t=1,2,\cdots,T. \tag{9}$$

The bias-removed signal $\hat{x}_{k,t}$ is passed through a second pass of Viterbi decoding procedure to reevaluate the likelihood score based on the word model $\Lambda(k)$. The mean bias which corresponds to the highest likelihood score is considered as the most probable mean bias $\hat{b}$.

$$\hat{b} = \arg\max_{\bar{b}_k}(\prod_t \max_i p(y_t - \bar{b}_k | \Lambda_i(k))). \tag{10}$$

Finally, the adjusted observation $\tilde{x}_t$ derived from the MSBR method is expressed as

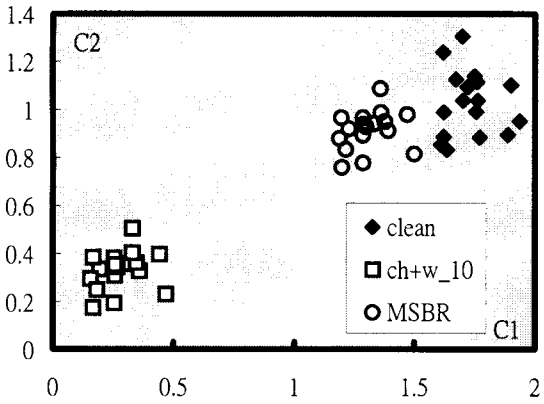
$$\tilde{x}_t = y_t - b \qquad t = 1,2,\cdots,T. \tag{11}$$

That is, the likelihood maximization procedure for bias estmation is primarily constrained within a local range, i.e., a word model, whereas the SBR is based on a global range, i.e., the set of all word models. The validation and effectiveness of above idea is illustrated in Fig. 3. From the scattering plots, we can observe that the distances between the clean cepstral clusters and the cepstral clusters compensated by the MSBR method are significantly reduced.

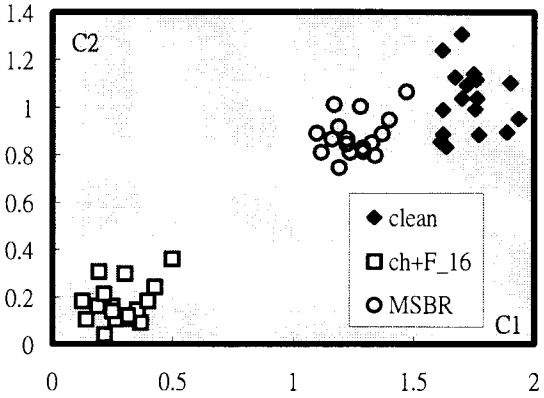
4. Experimental Results and Discussions

A multispeaker (50 males and 50 females) isolated Mandarin digit recognition was conducted to compare the effectiveness of the two blind channel equalization methods. There were three sessions of data collection and for each session every speaker uttered a set of 10 Mandarin digits. The first two sessions are used for training the word models and the other for testing. Each digit is modeled as a left-to-right HMM in which the output of each state is the mixture of two Gaussian distributions of feature vectors. Each feature vector consists of 12 LPC derived cepstral coefficients, 12 delta cepstral coefficients, and one delta log-energy. The additive ambient noises, including white noise and F16 colored noise, were added to clean speech with predetermined SNRs to generate various noisy speech signals. The channel effects were simulated by using 41 simulated telephone channel filters. Each testing utterance distorted by additive ambient noises is passed through a randomly selected channel filter to simulate the influence of channel effect.

The experiments are conducted on the following conditions : (1) with (denoted as HMM/Y) or without (denoted as HMM/N) iterative bias removal procedure in the training phase, (2) before passing through a channel filter, the clean speech signal is first corrupted by additive noises (specified by '0'-'20' dB) or not (specified by 'clean'). Two sets of experimental results are listed in Table 1. Table 1 (a) compares the performances of the three channel equalization methods and the baseline system in the case that white Gaussian noise is considered as an additive noise. In addition, the performance evaluation for the case of F16 colored noise is also shown in Table 1 (b). For all the results reported in Fig.1, the bias removal procedure is repeated three times. From those experimental results, we can see the following facts : (1) Comparing with the baseline system,



(a) for white noise with SNR at 10 dB.



(b) for F16 colored noise with SNR at I0 dB.

Fig. 3 The movement of cepstral clusters under various conditions.

the SBR method seems not so effective 5 when additive noises are taken into account. (2) The HML method is effective in compensating the speech signals transmitted over a telephone channel no matter the additive noises are considered or not. (3) The MSBR method outperforms the HML method and their performance are very close for low SNR.

5. Conclusions

In this paper, we first demonstrate that the SBR method is not so effective when additive noises are considered in telephone speech recognition. Then, we proposed a new method called MSBR which enables us to make further improvement in recognition accuracy. The MSBR method tries to select a more reliable and probable bias estimator among those estimators which are based on a local maximum likelihood estimation. Experimental results show that the MSBR is more effective in compensating the channel effect in telephone speech recognition.

Table 1. Comparisons of recognition rates

(a) for additive white noise

method	model	clean	20dB	15dB	10dB	5dB	0dB
BASELINE	HMM/N	93.9	63.5	45.8	28.1	16.5	11.4
SBR	HMM/N	95.0	64.5	47.5	30.0	17.3	11.4
HML	HMM/N	95.2	79.4	70.7	52.9	29.3	15.1
MSBR	HMM/N	95.7	81.7	72.5	54.3	29.8	15.2
SBR	HMM/Y	94.6	64.1	47.3	29.9	18.2	12.3
HML	HMM/Y	96.1	75.6	63.8	46.0	26.3	16.1
MSBR	HMM/Y	96.7	78.1	65.9	47.8	27.0	16.3

(b) for additive F16 colored noise

method	model	20dB	15dB	10dB	5dB	0dB
BASELINE	HMM/N	72.7	61.0	44.0	28.9	20.6
SBR	HMM/N	76.5	65.1	46.8	29.2	20.6
HML	HMM/N	87.9	81.1	67.0	41.1	21.2
MSBR	HMM/N	90.2	82.6	67.6	41.3	21.2
SBR	HMM/Y	76.8	64.3	46.1	25.9	18.3
HML	HMM/Y	82.4	73.1	55.1	33.7	19.3
MSBR	HMM/Y	84.8	74.8	56.0	34.0	19.8

References

[1] Mokbel, C., Paches-leal, P., Jouvet, 5D. and Monne, J. ""Compensation of telephone line effect for robust speech recognition"", Int. Conf. Spoken Language Processing, pp. 987-990, 1994.

[2] Hermansky, H. and Morgan, N. ""RASTA processing of speech"", IEEE Trans. Speech and Audio Processing, vol. 2, no. 4, pp. 578-589, 1994.

[3] Liu, F. H., Stern, R. M. Acero, A., and Moreno, P. J. ""Environmental normalization for robust speech recognition using direct cepstral comparison"", IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 1994, Vol. 2, pp. 61-64.

[4] Chien, J. T., Wang, H. C. and Lee, L. M. ""Estimation of channel bias for telephone speech recognition"", Int. Conf. Spoken Language Processing, pp. 1840-1843, 1996.

[5] Rahim, M. G. and Juang, B. H. ""Signal bias removal method for robust telephone based speech recognition in adverse environments,"" IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 445-448, 1994.

[6] Rahim, M. G. and Juang, B. H. ""Signal bias removal by maximum likelihood estimation for robust telephone speech recognition,"" IEEE Trans. on Speech and Audio Processing , vol. 4, no.1, pp. 19-30, 1996.

[7] Rahim, M. G., Juang, B. H., Chou, W. and Buhrke, E. ""Signal conditioning techniques for robust speech recognition,"" IEEE Signal Processing Letters. Vol. 3, No. 4, pp. 107-109, April 1996.

[8] Vaseghi, S. and Milner, B. ""A comparitive Analysis of channel-robust features and channel equalization methods for speech recognition, "" Int. Conf. Spoken Language Processing, pp. 877-880, 1996.