Joint Conference of 1996 International Computer Symposium
December 19~21, Kaohsiung, Taiwan, R.O.C.

An Extended Study on Signal Bias Removal 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

Abstract

Speech signals transmitted over telephone channel often suffer from interferences of additive ambient noise and multiplicative distortion. In this paper, a proportional alignment decoding (PAD) instead of Viterbi decoding (VD) is proposed and combined with a modified signal bias removal (MSBR) method to improve the performance and computation efficiency of a telephone speech recognition system. This proposed method simplifies the computation in decoding process and uses only the features in cepstral domain. Experiments on multispeaker (50 males and 50 females) isolated Mandarin digit recognition are conducted to demonstrate the effectiveness of the proposed method. Experimental results show that a remarkable improvement in recognition accuracy up to 20%-30% at 0 dB SNR can be achieved. Moreover, due to its simplicity in state alignment, the PAD method requires much less computation time in state decoding, and is more flexible in real-time applications.

1. Introduction

Telephone network plays an important role of information dissemination in our life. When an HMM-based speech recognition system is used in such an environment, undesired effects due to additive noise and multiplicative distortion will deteriorate the recognition accuracy and cause the recognizer unusable for real-world applications. The additive noise, such as ambient noises, has the effect to add an unflatten spectrum to the power spectrum of a clean speech [1]. This would result in the reduction of signal-to-noise ratio (SNR) and make the discrimination between different speech utterances hard. Another interference source, such as filtering effect of telephone channel and handset, affects speech by multiplying the spectrums of distortion sources to the spectrum of a clean speech.

In the implementation of a speech recognition system over telephone network, it is known as shown in Fig. 1 that the speech signal s(t) mixed with the additive ambient noise n(t) is transmitted over the telephone line to produce a distorted signal y(t) at the speech recognition system input. A telephone network generally behaves as a linear convolved filter h(t). The distorted signal y(t) must be compensated at the time of speech recognition process. Many researches dealt with only one of the interference sources. Mokbel et al. [2]-[3] assumed that the speakers talk close to the telephone handset and used a cepstral mean subtraction (CMS) method to eliminate the channel effect of a telephone line. In this method, the averaged cepstral coefficient vector of the whole utterance is subtracted from each frame of the speech utterance. Hermansky et al. [4]-[5] proposed the use of RelAtive SpecTrAl (RASTA) for the compensation of telephone channel effect. Their approach performed band-pass filtering on the log-spectrum domain to reduce the convolved components in the contaminated speech signals based on the characteristics of human perceptive system. In the papers [6]-[7] proposed by Rahim and Juang, they dealt with the undesired signal components due to additive ambient noise and channel distortion as "signal biases". A signal bias removal (SBR) method based on maximum likelihood estimation is presented for the minimization of these extraneous components. Theoretically, the SBR method can be iterated between spectral domain (for additive ambient noise) and cepstral domain (for convolved noise) until extraneous effects are minimized. However, the results presented in the papers [6]-[7] only consider a multiplicative spectral bias.

In this paper, an efficient method which incorporates a proportional alignment decoding algorithm (PAD) [8] with a modified SBR method (MSBR) is proposed to improve the performance of a speech recognition system over telephone line. When the multiplicative and additive spectral biases are considered in implementing a telephone speech recognition system, our approach performs training and recognition procedures only in cepstral domain whereas the SBR method have to transform speech signals between spectral and cepstral domains. In addition, the PAD method is also much more computationally efficient than Viterbi decoding method used in SBR for state decoding.

This paper is organized as follows. In the next section, we briefly describe the formulation of the signal bias removal method. And then, a modified signal bias removal method used in this paper and a proportional alignment decoding method are presented. Subsequently, the procedures for applying the MSBR and PAD in both training and recognition phases of a telephone speech recognition system are discussed. In section 3, the underlying speech database, some related testing conditions, and experimental results are illustrated. Finally, a conclusion is drawn in Section 4.

2. Compensation for Telephone Speech Recognition

2.1 Formulation of Signal Bias Removal Method (SBR) [6]-[7]

As shown in Fig. 1, let y(t) be the received contaminated signal, s(t) the original clean signal, n(t) the additive noise signal and h(t) the linear transfer function of a telephone line. Then the distorted signal y(t) can be represented as [2]

$$y(t) = \left[s(t) + n(t)\right] \otimes h(t), \qquad (1)$$

where $\otimes$ denotes the convolution operation. By using Fourier transform and taking the logarithm on Eqn. (1), the logarithmic spectrum $Y(f)$ can be modeled by

$$\begin{aligned} \log|Y(f)| &= \log|S(f)+N(f)| + \log|H(f)| \\ &= \log|S(f)| + \log\left|1+\frac{N(f)}{S(f)}\right| + \log|H(f)|, \end{aligned} \qquad (2)$$

where $Y(f)$, $S(f)$, $N(f)$ and $H(f)$ are the spectra of *y(t), s(t), n(t)* and *h(t)*, respectively. Applying the inverse Fourier transform, we can project Eqn. (2) on the cepstral domain and yields

$$C_y(\tau) = C_s(\tau) + C(\tau) + C_h(\tau), \qquad (3)$$

where $C_y(\tau)$, $C_s(\tau)$, $C(\tau)$ and $C_h(\tau)$ are the cepstra derived from spectra *Y(f), S(f),* $1+{}^{N(f)}\!/_{S(f)}$ and *H(f)*, respectively. In Eqns. (2) and (3), *N(f)* and *H(f)* are considered as "biases" in spectral domain while $C(\tau)$ and $C_h(\tau)$ in cepstral domain. The SBR procedure is used to estimate these biases in both domains and try to minimize their effects on the speech signals.

For completeness, the SBR method proposed by Rahim and Juang is briefly described as follows. This method is based on maximization of the likelihood defined as

$$p(X|\Lambda) = \max_{\Lambda(w)} p(X|\Lambda(w)) \qquad (4)$$

and

$$p(X|\Lambda(w)) = \prod_t \max_i p(x_t|\Lambda_i(w)), \qquad (5)$$

where $X=\{x_1,x_2,\cdots,x_t,\cdots,x_T\}$ is an observation sequence of *T* frames. The set of all word models is represented by $\Lambda = \{\Lambda(w),\ w = 1,2,\ldots,N\}$ and $\Lambda(w) = \{\Lambda_i(w),$ $i = 1,2,\ldots,M_w\}$, where N is the number of word models, M_w is the number of states in the *w*-th word model $\Lambda(w)$, and $\Lambda_i(w) = N(\mu_i(w),\Sigma_i(w))$ is a

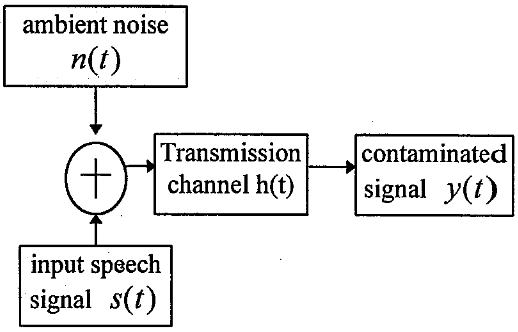

Fig.1 block diagram for telephone speech recognition

Gaussian probability with the mean vector and covariance matrix of *i*-th state in the *w*-th word model. When an additive bias term *b* is considered, the contaminated signal $Y=\{y_1,y_2,\cdots,y_t,\cdots,y_T\}$ can be expressed as

$$y_t = x_t + b \qquad t = 1,2,\cdots,T. \qquad (6)$$

Then the maximum likelihood bias estimator $\bar{b}$, based on the set of all word models Λ, can be formulated as

$$\bar{b} = \frac{1}{T}\sum_{t=1}^{T}(y_t - \delta_t), \qquad (7)$$

where

$$\begin{aligned} \delta_t &= \arg\max_{\Lambda_i(w^*)} p(y_t|b,\Lambda_i(w^*)) \\ &= \arg\max_{\Lambda_i(w^*)} p(y_t - b|\Lambda_i(w^*)), \end{aligned} \qquad (8)$$

and

$$w^* = \arg\max_w p(Y|b,\Lambda(w)). \qquad (9)$$

2.2 A Modified Signal Bias Removal Method (MSBR)

In our approach as shown in Fig. 2, the contaminated signal Y is evaluated on every word model to estimate the mean biases corresponding to different hidden Markov models, i.e.,

$$\bar{b}_k = \frac{1}{T}\sum_{t=1}^{T}(y_t - \delta_{k,t}) \qquad k = 1,2,\cdots,N, \qquad (10)$$

where

$$\begin{aligned} \delta_{k,t} &= \arg\max_{\Lambda_i(k)} p(y_t|b_k,\Lambda_i(k)) \\ &= \arg\max_{\Lambda_i(k)} p(y_t - b_k|\Lambda_i(k)). \end{aligned} \qquad (11)$$

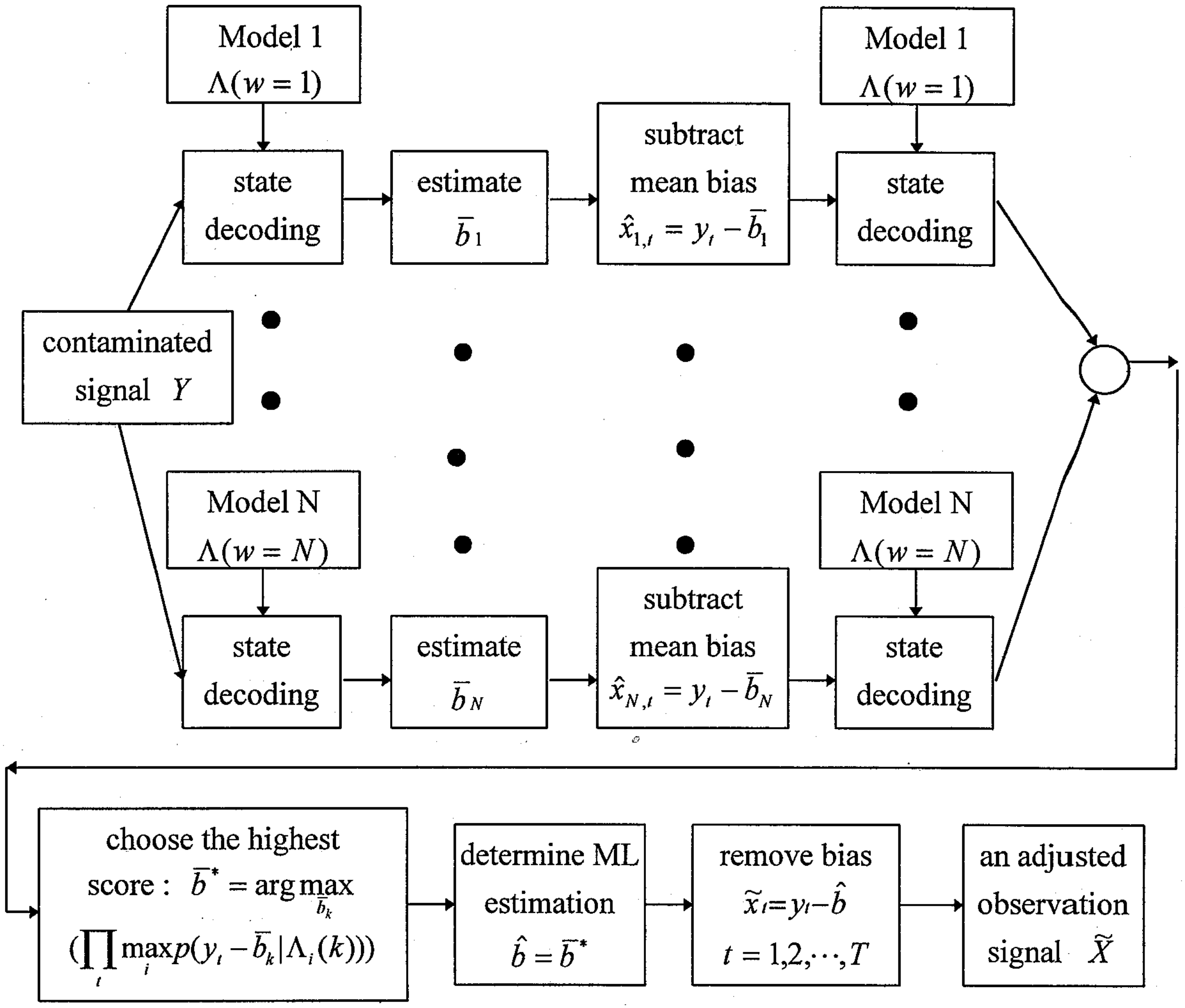
Model 1
$\Lambda(w=1)$
state decoding
estimate $\bar{b}_1$
subtract mean bias $\hat{x}_{1,t}=y_t-\bar{b}_1$
Model 1
$\Lambda(w=1)$
state decoding
contaminated signal Y
Model N
$\Lambda(w=N)$
state decoding
estimate $\bar{b}_N$
subtract mean bias $\hat{x}_{N,t}=y_t-\bar{b}_N$
Model N
$\Lambda(w=N)$
state decoding
choose the highest score : $\bar{b}^*=\arg\max_{\bar{b}_k}(\prod_t \max_i p(y_t-\bar{b}_k|\Lambda_i(k)))$
determine ML estimation $\hat{b}=\bar{b}^*$
remove bias $\tilde{x}_t=y_t-\hat{b}$ $t=1,2,\cdots,T$
an adjusted observation signal $\tilde{X}$

Fig.2 a modified signal bias removal (MSBR) method

Once a mean bias $\bar{b}_k$ is obtained for the word model $\Lambda(k)$, this mean bias is subtracted from the original contaminated signal to get a bias-removed signal $\hat{x}_{k,t}$,

$$\hat{x}_{k,t}=y_t-\bar{b}_k \quad k=1,2,\cdots,N, \quad t=1,2,\cdots,T. \tag{12}$$

A second pass of state decoding procedure is applied to the bias-removed signal $\hat{x}_{k,t}$ to reevaluate the likelihood score based on the word model $\Lambda(k)$. The mean bias corresponding to the highest likelihood score is used as the maximum likelihood estimator $\hat{b}$ of the extraneous bias b, that is

$$\hat{b}=\arg\max_{\bar{b}_k}(\prod_t \max_i p(y_t-\bar{b}_k|\Lambda_i(k))). \tag{13}$$

Then the adjusted observation $\tilde{x}_t$ after bias removal is

$$\tilde{x}_t=y_t-\hat{b} \qquad t=1,2,\cdots,T, \tag{14}$$

and

$$\tilde{X}=\{\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_t,\cdots,\tilde{x}_T\}. \tag{15}$$

2.3 A Proportional Alignment Decoding Algorithm (PAD) [8]

In order to reduce the computation complexity in state decoding process, a proportional alignment decoding method instead of the widely used Viterbi decoding is proposed. This method can be applied in both training and recognition phases of a conventional HMM-based speech recognition system. At first, the utterances from the training database are used to obtain the word models. By using these word models and the Viterbi decoding algorithm to all the training utterances, we can decode each training utterance into state sequence.

2.3.1 **Calculation of State Duration Ratio**

By using the state sequences decoded for training utterances, we can calculate the state duration mean for each state. For the case of state i in word w, its state duration mean is

$$\bar{d}(i,w)=\frac{1}{N_w}\sum_{j=1}^{N_w} d(i,w,j), \qquad (16)$$

where N_w is the number of training utterances for word w, and $d(i,w,j)$ is the duration of state i in word w for j-th training utterance. The state duration is expressed in terms of number of frames. The word duration mean is the accumulation of all the state duration means in a word.

$$\bar{d}(w)=\sum_{i=1}^{S_w}\bar{d}(i,w), \qquad (17)$$

where S_w is the number of states in the hidden Markov model of word w. Then the duration ratio of state i in word w is calculated by

$$r(i,w)=\frac{\bar{d}(i,w)}{\bar{d}(w)} \qquad (18)$$

2.3.2 **State Decoding Process of PAD**

Once we obtain the state duration ratio $r(i,w)$ for all word models, the state decoding process can be proceeded in a direct way. For example, an utterance j of word w has duration of $d(w,j)$ frames. We align the duration for each state in this utterance by the following equation

$$\tilde{d}(i,w,j)=\lceil r(i,w)\times d(w,j)-0.5\rceil \;, \qquad (19)$$

where $\lceil x \rceil$ denotes the smallest integer which is greater than or equal to x. From above description, it is known that the PAD method can prevent any state from staying too long or too short in decoding a state sequence. This leads to a more accurate match of state durations between distorted speech and reference models. Also, due to its simple way in state decoding, the proposed decoding method requires much less computation cost and complexity in state decoding.

2.3.3 **MSBR & PAD in Training Phase**

During the training phase of our approach, the utterances from the training database are first used to create those initial word models by means of the generalized forward-backward reestimation method and the Viterbi decoding algorithm. The duration of each state in every training utterance is realigned with the corresponding initial word model by using the PAD method so that the mean bias is estimated. Once the mean bias is subtracted from each frame of the training utterance, we recalculate the feature vectors of each state for every word model to come out a new set of reference models. By using these new models and the PAD method, we decode the state sequences of those bias-removed training utterances again. Finally, an adjusted reference model is obtained.

2.3.4 **MSBR & PAD in Recognition Phase**

In the recognition phase, we first use the MSBR method employing the PAD method for state decoding to remove the bias from the contaminated testing utterance Y. Then the adjusted observation signal $\tilde{X}$ is evaluated on every adjusted reference model and decoded by using the PAD method to find an optimum state sequence, and by which the likelihood score is calculated. Finally, the most probable word model corresponding to the testing utterance can be obtained. It is worth to note that all of the state decoding operations in this phase is based on the

Table 1. recognition rates excluding channel effect
(a) for additive white noise

train phase	recog phase	clean	20dB	15dB	10dB	5dB	0dB
VD	VD	99.0	81.2	65.9	48.9	25.3	10.4
VS	VS	98.1	81.9	65.7	46.5	24.1	10.4
VM	VM	97.2	82.6	67.4	47.9	25.3	16.7
PM	PM	95.3	87.4	81.0	72.6	56.6	41.0

1. VD : Viterbi decoding algorithm
2. VS : VD + signal bias removal method
3. VM :VD + modified signal bias removal method
4. PAD : proportional alignment decoding method
5. PM: PAD + modified signal bias removal method

(b) for additive F16 colored noise

train phase	recog phase	20dB	15dB	10dB	5dB	0dB
VD	VD	92.4	78.1	65.2	43.6	22.2
VS	VS	91.3	78.8	64.1	40.6	21.2
VM	VM	90.6	81.9	65.8	41.8	22.8
PM	PM	92.6	89.9	81.6	65.5	47.2

Table 2. recognition rates including channel effect
(a) for additive white noise

train phase	recog phase	clean	20dB	15dB	10dB	5dB	0dB
VD	VD	93.5	63.5	45.8	28.1	16.5	11.4
VS	VS	94.5	63.7	46.5	24.9	16.9	11.4
VM	VM	95.4	65.7	47.4	26.8	17.6	16.2
PM	PM	95.0	86.7	80.1	71.5	55.3	36.9

(b) for additive F16 colored noise

train phase	recog phase	20dB	15dB	10dB	5dB	0dB
VD	VD	72.7	61.0	44.0	28.9	20.6
VS	VS	75.5	62.4	43.8	24.7	18.2
VM	VM	78.4	64.1	44.7	25.9	19.6
PM	PM	89.9	87.4	81.0	62.0	43.6

PAD method instead of Viterbi decoding method. Besides, we use the MSBR method once only for bias removal in contrast to the SBR method in iterative manner.

3. Experimental Results and Discussions

A multispeaker (50 males and 50 females) isolated Mandarin digit recognition was conducted to demonstrate the effectiveness of the proposed method. The speech signal is sampled at 8 KHz. Each frame contains 256 samples with 128 samples overlapped, and is multiplied by a 256-point Hamming window. There are 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 at infinity, 20, 15, 10, 5 and 0 dB to generate various noisy speech signals. The channel filtering effects were simulated by using 41 simulated telephone channel filters. These filters are represented by 49-order AR models, and the filter parameters were determined by optimally fitting the frequency responses of various telephone handsets. Each testing utterance distorted by additive ambient noises is passed through a randomly selected channel filter to simulate the influence of channel filtering effect. In our experiments, a conventional continuous-density hidden Markov model is referred to as a baseline system.

The experiments are arranged based on following considerations : (1) MSBR using Viterbi decoding or proportional alignment decoding for retraining word models in the training phase, (2) MSBR using Viterbi decoding (VD+MSBR) or proportional alignment decoding (PAD+MSBR) in the recognition phase. For comparative purpose, the SBR method is also implemented. During the training phase, the procedure of estimating and removing the bias is repeated two times in cepstral domain only. All the feature vectors are then recomputed to generate a new set of reference models. In the recognition phase, an estimate of the bias is calculated for each testing utterance and subtract from it. Similarly, this procedure is also repeated two times in cepstral domain.

Two sets of experimental results are listed in Table 1 and Table 2. One is for the case excluding channel filtering effect, and the other is for the case including channel filtering effect. From the experimental results, we can see the following facts : (1) When channel filtering effect is considered in the environment with additive noises, the recognition performance will degrade severely. (2) The recognition accuracy of the MSBR method is slightly better than that of the SBR method. (3) The SBR and MSBR methods can reduce the influence of channel filtering effect. However, when additive noises are also considered, they are no longer effective for telephone speech recognition. (4) For the case excluding channel filtering effect, the (PAD+MSBR) improve recognition accuracy from 10%~22% to 41%~47% at 0dB as compared with the baseline and (VD+MSBR) for various additive noise sources. (5) When channel filtering effect is introduced, the (PAD+MSBR) still makes a significant improvement in recognition rates. In some case, the performance is even better. (6) The PAD method is effective in reducing the influence of additive ambient noises. This is because that the PAD method can model the temporal structures of speech utterances in a reasonable manner and assure a more reliable recognition process in noisy environment. Finally, it is worth to mention that the PAD method decodes state sequence in a direct alignment manner and thus is much more computationally efficient in recognition procedure.

4. Conclusions

In this paper, a proportional alignment decoding method combined with a modified signal bias removal method is successfully applied in telephone speech recognition when the multiplicative and additive spectral biases are considered. Experimental results show that the proposed method can provide remarkable improvement in speech recognition than those of a conventional HMM when the speech is transmitted over telephone network with different noisy conditions. In addition, the MSBR method is superior slightly to the SBR method for estimation of channel distortion. Some important features of our approach are summaried as following : (1) The proposed method is effective to reduce the influences due to additive noise and multiplicative distortion. (2) The PAD method can be applied individually to additive noises or combined with MSBR to additive and multiplicative noises. (3) The training and recognition procedures of our approach are performed in cepstral domain only, no additional transformation between cepstral and spectral domains is needed. (4) Due to its simplicity in state alignment, the PAD method requires much less computation time and complexity in state decoding. This feature will meet the requirement of real-time speech recognition.

References

[1] Rabiner, L. and Juang, B. H. : Fundamentals of speech recognition, Englewod Cliffs, N.J.:Prentice Hall, 1993.

[2] Mokbel, C., Monne, J. and Jouvet, D. “”On-line adaptation of a speech recognizer to variations in telephone line conditions“”, European Conference on

Speech Communication and Technology (EUROSPEECH), pp. 1247-1250, 1993.

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

[4] Hermansky, H., Morgan, N., Bayya, A. and Kohn, P. “”Compensation of the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)“”, European Conference on Speech Communication and Technology (EUROSPEECH), pp. 1367-1370, 1991.

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

[6] 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.

[7] 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.

[8] Hung, W. W. and Wang, H. C. “”Noisy speech recognition based on state duration alignment,“” submitted for publication in IEE Proceedings. Vision, Image and Signal Processing.