ESTIMATION OF CHANNEL BIAS FOR TELEPHONE SPEECH RECOGNITION

*Jen-Tzung Chien, Hsiao-Chuan Wang and Lee-Min Lee**

Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan
*Department of Electrical Engineering, Mingchi Institute of Technology, Taipei Hsien, Taiwan

ABSTRACT

In this study, we propose a maximum a posterior (MAP) estimation of channel bias to compensate the channel mismatch in telephone speech recognition. For a telephone speech, the channel bias is estimated by maximizing a posterior probability. Because a posterior probability is composed of a likelihood function and a prior density, we introduce a scale factor to evaluate their weights in MAP estimation. To further improve the performance, a prior channel statistics is extended to multiple components and the channel mismatch is separately compensated for different segments. Besides, a rapid MAP estimation applied in feature domain is also proposed for reducing the computational complexity. Experiments show that proposed method can significantly improve recognition rates and computational complexity.

1. INTRODUCTION

In real application, the automatic speech recognition system is often applied in adverse environment. The speech recognition over telephone network is such a widespread application. Basically, the distortion sources in telephone network come from two classes: (1) noise contamination including background noise and electrical noise, and (2) channel effect caused by telephone handset and transmission line. Due to these distortion sources, the speech recognition performance will be seriously damaged. In the literature, many algorithms [1] have been proposed for compensating the noise effect. However, it is not adequate for overcoming the mismatch problem by only considering the noise effect. Accordingly, the codeword-dependent cepstral normalization (CDCN) [2] and RASTA method [3] were presented for reducing the variability of additive noise and channel effect. Besides, the signal bias removal (SBR) [4], stochastic matching (SM) [5] and channel-effect-cancellation [6] methods were also successfully applied for telephone speech recognition.

This paper propose the MAP estimation of channel bias between telephone speech and reference models [7]. The channel bias is estimated by maximizing a posterior probability of channel bias given the decoded state sequence. This method can be applied in model space for adapting reference models as well as in the feature space for canceling the channel effect. For model adaptation, a channel bias between the telephone speech and a reference pattern is estimated by MAP estimation. Each reference pattern is adapted by its corresponding channel bias. On the other hand, for channel cancellation, the telephone speech is enhanced by canceling the channel bias which is obtained by MAP estimation based on a well-aligned state sequence. In this study, a scale factor is introduced to MAP estimation for assessing the weights of a likelihood function and a prior probability. A prior statistics of channel bias is also extended to multiple components such that a closer prior statistics can be merged in MAP estimation. In addition, a rapid MAP estimation scheme is proposed for effectively reducing the computational cost. The experiments for recognizing 250 Mandarin names show that proposed method can efficiently overcome channel mismatch in telephone speech recognition.

2. MAP CHANNEL ESTIMATION

The formula derivation of MAP channel estimation is described as follows: Let $\mathbf{h}$ denote the channel bias between the observation sequence $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_T\}$ and a reference pattern given the decoded state sequence $\mathbf{S} = \{s_1, s_2, \cdots, s_T\}$. Here, the feature vectors and the channel bias are represented in cepstrum. The channel bias $\mathbf{h}$ is characterized by a multivariate Gaussian pdf with mean vector μ_h and covariance matrix Σ_h. The state parameter of hidden Markov models (HMMs) s_t is also a multivariate Gaussian pdf $N(\mu_{s_t}, \Sigma_{s_t})$. Based on MAP criterion, the channel bias $\mathbf{h}$ is estimated by maximizing a posterior probability $P(\mathbf{h}|\mathbf{Y},\mathbf{S})$, i.e.

$$\hat{\mathbf{h}}_{MAP} = \arg\max_{\mathbf{h}} P(\mathbf{h}|\mathbf{Y},\mathbf{S}) \tag{1}$$

This equation is also equivalent to

$$\hat{\mathbf{h}}_{MAP} = \arg\max_{\mathbf{h}} \{\log P(\mathbf{Y}|\mathbf{h},\mathbf{S}) + \log P(\mathbf{h})\} \tag{2}$$

That is, the maximization problem is transferred to maximize the sum of log likelihood $\log P(\mathbf{Y}|\mathbf{h},\mathbf{S})$ and logarithm of a prior pdf $\log P(\mathbf{h})$. This motivates us to introduce a scale factor α into Eq. (2) for evaluating the weights of these two terms. Thus, we generalize the MAP estimation as follows.

$$\hat{\mathbf{h}}_{MAP} = \arg\max_{\mathbf{h}} \{\alpha \log P(\mathbf{Y}|\mathbf{h},\mathbf{S}) + (1-\alpha) \log P(\mathbf{h})\} \tag{3}$$

By extending these two terms in frame sequence, we can find

$$\hat{\mathbf{h}}_{MAP} = \left(\alpha \sum_{t=1}^{T} \Sigma_{s_t}^{-1} + (1-\alpha) T \Sigma_h^{-1} \right)^{-1} \cdot \left(\alpha \sum_{t=1}^{T} \Sigma_{s_t}^{-1} (\mathbf{y}_t - \mu_{s_t}) + (1-\alpha) T \Sigma_h^{-1} \mu_h \right) \quad (4)$$

If the scale factor $\alpha = 1$, the equation is reduced to the maximum likelihood (ML) estimation, i.e.

$$\hat{\mathbf{h}}_{ML} = \left(\sum_{t=1}^{T} \Sigma_{s_t}^{-1} \right)^{-1} \left(\sum_{t=1}^{T} \Sigma_{s_t}^{-1} (\mathbf{y}_t - \mu_{s_t}) \right) \quad (5)$$

Observing from Eq.(4), we find that the MAP channel estimator is determined by two factors. One is the cepstral average of difference of observation vectors and corresponding HMM mean vectors. The other one is a prior channel mean vector. They are weighted by the covariance matrices Σ_{s_t} and Σ_h respectively and interpolated by a scale factor α. If a prior channel statistics is properly chosen, the estimation error of MAP channel estimator caused by incorrect state sequence can be adequately compensated. Thus, a prior channel statistics play an important role in MAP channel estimation.

2.1 Extension of a prior channel statistics

In general, a well-defined prior channel statistics should be extracted by sufficient training data which covers all the variabilities of channel characteristics. The resulting MAP channel estimation will be more reliable. However, a single prior channel statistics is not suitable for each MAP channel estimation. It is because that a prior channel statistics may not be close to the channel bias of telephone speech. In this case, a prior channel statistics can not effectively compensate the estimation error. To increase the correctness of a prior channel statistics in MAP channel estimator, we are motivated to cluster the training data of channel bias into multiple codebooks using vector quantization. Then, a set of a prior channel statistics $\{N(\mu_h^c, \Sigma_h^c), 1 \le c \le C\}$ is generated. The channel bias of telephone speech is characterized by this set of a prior statistics. When a prior channel statistics with multiple components is utilized, the MAP channel estimator is modified by applying the following procedure :

1. Use Eq. (5) for estimating ML channel bias $\hat{\mathbf{h}}_{ML}$.
2. According to ML channel bias $\hat{\mathbf{h}}_{ML}$, the closest codebook among the set of a prior channel statistics is extracted by

$$\hat{c} = \arg\max_{c} P(\hat{\mathbf{h}}_{ML} | c) \quad (6)$$

3. Substitute corresponding mean vector and covariance matrix into Eq. (4). A modified MAP channel estimator is shown by

$$\hat{\mathbf{h}}_{MAP} = \left(\alpha \sum_{t=1}^{T} \Sigma_{s_t}^{-1} + (1-\alpha) T \Sigma_h^{\hat{c}-1} \right)^{-1} \cdot \left(\alpha \sum_{t=1}^{T} \Sigma_{s_t}^{-1} (\mathbf{y}_t - \mu_{s_t}) + (1-\alpha) T \Sigma_h^{\hat{c}-1} \mu_h^{\hat{c}} \right) \quad (7)$$

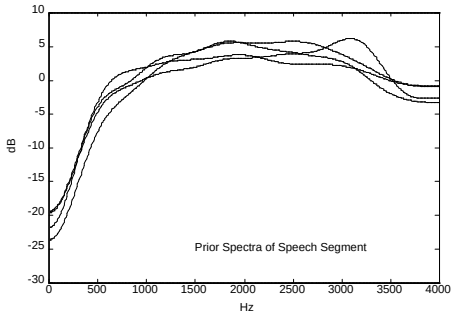
By incorporating the closest prior channel statistics, the estimation correctness can be increased.

2.2 Compensation for different segments

In general, a telephone utterance can be divided into speech segment and non-speech segment. For speech segment, if the signal-to-noise (SNR) of telephone speech is large, the distortion source is dominated by the channel effect. If SNR is small, the speech segment is distorted by various degrees of noise effect and channel effect. However, for non-speech segment, no matter what SNR is, the noise effect is always existed. Thus, we conclude that the main distortion sources of speech segment and non-speech segment are different [5]. To reflect this phenomenon, the channel biases of speech segment and non-speech segment should be separately estimated. Also, the frames of speech segment and non-speech segment should be separately compensated. In this study, the MAP channel estimation is applied for estimating these two channel biases. The corresponding prior statistics of different segments is also reestimated according to the frame sequence of different segments.

3. A PRIOR CHANNEL STATISTICS

A prior channel statistics is an important part of MAP channel estimation. To effectively estimate a prior channel statistics containing sufficient channel characteristics, we collect 8278 telephone utterances and respectively estimate their corresponding channel biases. The estimation of channel bias vectors are supervised by using iterative ML channel estimator [7]. Here, the reference models are trained from 5045 syllable-balanced words (51 males and 50 males) recorded by a high-quality microphone. When a prior channel statistics is separately estimated for different segments, the spectra of mean vectors are shown in Fig.1.



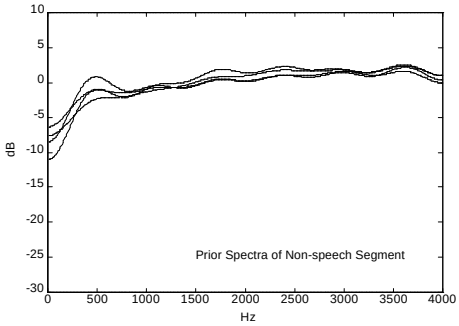


Figure 1: Spectra of 4 codebooks of a prior channel statistics for different segments.

4. MAP CHANNEL ESTIMATION IN FEATURE SPEACE

In previous study [7], the proposed MAP channel estimation is employed in model adaptation. For model adaptation, because each reference pattern is adapted by its corresponding channel bias estimated by a decoded state sequence, the computational complexity will be greatly increased for a large-vocabulary recognition task. Actually, the MAP channel estimation can be also applied in feature space. That is, if we have a well-aligned state sequence for a telephone speech, the corresponding channel bias can be accurately estimated. By applying the cepstral subtraction, the telephone speech can be enhanced. The flow chart of MAP channel estimation in feature space is shown in Fig. 2. The channel bias of telephone speech is estimated according to a reliable state sequence of the highest score during the first pass Viterbi decoding. At the same time, the boundary of speech segment and non-speech segment is detected. After the channel bias is estimated by MAP channel estimator, the telephone speech can be enhanced by subtracting corresponding channel bias. In our method, since a prior channel statistics is employed in the first pass Viterbi decoder, the additional operation for iteratively updating the enhanced speech is negligible. Thus, by using this method, we can expect that the computational load can be largely reduced.

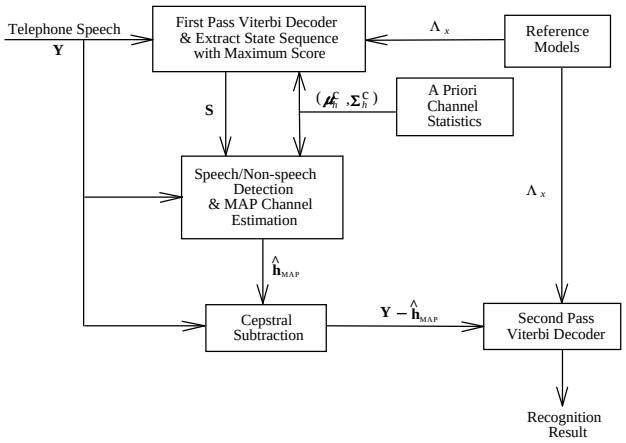


Figure 2: Flow chart of MAP channel estimator in feature space.

5. RAPID MAP CHANNEL ESTIMATION

Although the computational amount is significantly reduced by applying the MAP channel estimator in feature space, the real-time implementation is still not easily achieved. From our observation, the computational amount of MAP channel estimation is almost spent in the first pass Viterbi decoder for finding the closest reference pattern and its corresponding state sequence. To further decrease the computational amount, a set of segment-based HMM parameters are introduced in MAP channel estimation.

Because the Mandarin syllable is decomposed into an initial/final format similar to the consonant/vowel relation in other languages, we arrange 3 and 4 HMM states for characterizing initial and final parts of an Mandarin syllable. That is, a Mandarin syllable can be composed of 7 phonetic segments $\{i_1, i_2, i_3, f_1, f_2, f_3, f_4\}$. Thus, we are motivated to generate a new set of segment-based HMM parameters $\Lambda_{seg} = \{\lambda_{i_1}, \lambda_{i_2}, \lambda_{i_3}, \lambda_{f_1}, \lambda_{f_2}, \lambda_{f_3}, \lambda_{f_4}, \lambda_{sln}\}$ where λ_{sln} is the HMM parameter of silence. Each segment-based HMM parameter represents the phonetic characteristics of corresponding segment of 408 Mandarin syllables. The segment-based HMM parameters is generated as follows. First, by applying the segmental k-means algorithm, each training syllable is segmented into seven segments. Then, we collect the training frames of the same segment and apply vector quantization technique for generating the segment-based HMM parameters Λ_{seg}. To sufficiently reflect the segment-based phonetic characteristics, each segment-based HMM parameter is consisted of 16 mixture components. Based on the segment-based HMM parameters, no matter how large the vocabulary size is, the telephone speech can be rapidly time-aligned and corresponding channel bias can be efficiently calculated. After the feature enhancement and the second pass Viterbi decoding, the rapid MAP channel estimation is completed

6. EXPERIMENTAL RESULTS

A multispeaker (37 males and 36 females) task for recognizing 250 Mandarin names is conducted to demonstrate the proposed method. A total of 1000 utterances were collected via telephone network and recorded by using 10 telephone handsets. 12-order cepstral coefficients, 12-order delta cepstral coefficients, 1 delta log energy and 1 delta delta log energy are used for characterizing a speech frame. The recognition system is constructed by the continuous-density HMM. Each HMM state is modeled by 1-mixture Gaussian density. For comparative study, the cepstral mean normalization (CMN) is included in the experiments . As listed in Table 1, the recognition rates of baseline system and CMN method are 43.9% and 77.3% respectively. This reveals the effectiveness of CMN method. In the following, several sets of experiments will be reported. First, to evaluate the scale factor α and the codebook size of a prior channel statistics C, a set of experiments are shown in Fig. 3. We can see that the recognition rate can be improved to 77.6% when $C = 4$ and $\alpha = 0.9$. Thus, the extension of a prior channel statistics is preferable for MAP channel estimation. In the second set of experiments, we want to

evaluate the effectiveness of separate compensation of speech segment and non-speech segment. The results are illustrated in Fig. 4. We can see that the recognition rate is greatly improved to 85.3% by applying this technique. The best result is obtained when $\alpha = 0.6$. The previous two sets of experiments are reported when the MAP channel estimation is applied in model space. In the third set of experiments, the recognition rates and the computational speed are compared in Table 1. The computational speed (evaluated by seconds/utterance) is determined by simulating the proposed method on Sun SPARCstation 20 with model 50. Here, only results of $\alpha = 0.6$ are listed. We can see that the computational speed is effectively improved when the MAP channel estimator is applied in feature space. However, when the rapid MAP channel estimation is applied, the computational speed is almost comparable to that of baseline system. The recognition rates of three kinds of MAP channel estimation are superior to that of CMN method. From these results, we conclude that proposed method can rapidly and effectively overcome the channel mismatch in telephone speech recognition.

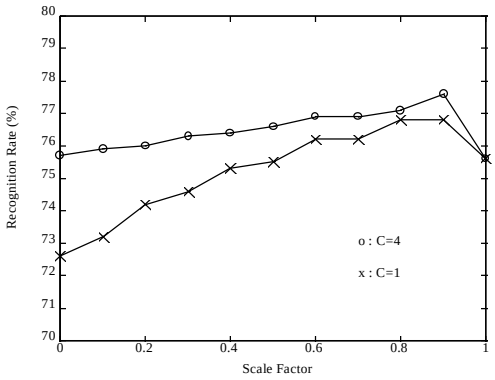


Figure 3: Recognition results of different scale factors and codebook sizes.

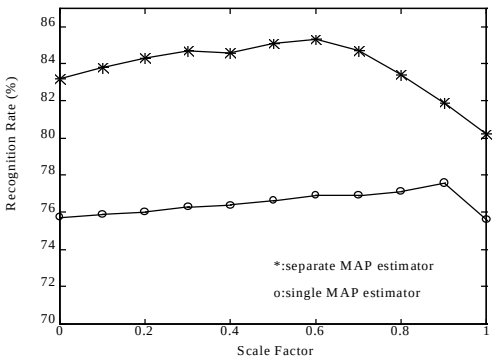


Figure 4: Recognition results of two kinds of MAP channel estimator.

	Baseline	CMN	MAP (Model)	MAP (Fea.)	Rapid MAP
Recog. (%)	43.9	77.3	85.3	83.8	82.6
Speed (s./utter.)	1.56	1.57	7.89	2.72	1.69

Table 1: Comparison of recognition rates and computational speed of baseline, CMN and three kinds of MAP channel estimatior.

7. CONCLUSION

This paper presents an estimation method of channel bias for telephone speech recognition. A MAP estimation of channel bias is derived for model adaptation as well as feature enhancement. To improve the performance, a prior channel statistics is extended to multiple components and the compensation of different segments are separately considered. To reduce computational overhead, a rapid MAP channel estimator is presented. From the recognition results, we find that the recognition rates and computational speed of MAP channel estimation are significantly improved. Therefore, we conclude that the proposed method is efficient for telephone speech recognition.

8. ACKNOWLEDGMENT

The authors acknowledge the support of Telecommunication Laboratory, MOCT, Taiwan, R.O.C., under contract TL-85-5203. We also thank useful discussions with E.F. Huang and C.S. Huang.

9. REFERENCES

1. Gong, Y. "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, pp. 261-291, 1995.
2. Acero, A. and Stern, R.M. "Environmental robustness in automatic speech recognition," in *Proc. ICASSP*, vol. 1, pp. 849-852, 1990.
3. Hermansky, H., Morgan, N., Bayya, A. and Kohn, P. " Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)," in *Proc. EUROSPEECH*, vol. 3, pp. 1367-1370, 1991.
4. Rahim, M.G. and Juang, B.H. "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, pp. 19-30, 1996.
5. Sankar, A. and Lee, C.H. "Robust speech recognition based on stochastic matching," in *Proc. ICASSP*, pp. 121-124, 1995.
6. Chien, J.T., Lee, L.M. and Wang, H.C. "A channel-effect-cancellation method for speech recognition over telephone system," *IEE Proc. Vis., Image and Signal Process.*, vol. 142, no. 6, pp. 395-399, 1995.
7. Chien, J.T., Lee, L.M. and Wang, H.C. "Channel estimation for reference model adaptation in telephone speech recognition," in *Proc. EUROSPEECH*, vol. 2, pp. 1541-1544, 1995.