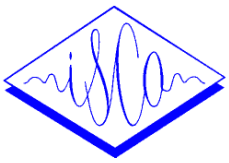

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

CHANNEL ESTIMATION FOR REFERENCE MODEL ADAPTATION IN TELEPHONE SPEECH RECOGNITION

Jen-Tzung Chien, Lee-Min Lee, and Hsiao-Chuan Wang
Department of Electrical Engineering,
National Tsing Hua University,
Hsinchu, Taiwan.

ABSTRACT

This paper deals with a channel estimation problem in speech recognition over telephone networks. In this study, we propose a channel estimation method to adapt the reference models so that the reference models can match the testing environment. For a telephone speech, its channel cepstral vector is estimated by the maximum a posteriori criterion. The reference models are then adapted to the testing environment by merging the estimated channel cepstral vectors. The telephone speech is recognized by using the adapted reference models. Experiments show that the proposed method can well estimate the telephone channel spectra and is successfully applied for telephone speech recognition.

1. INTRODUCTION

Speech recognition system is usually used in an adverse environment, such as over the telephone networks. The speech signal is distorted by additive noise and linear filtering effect. The mismatch between training and testing environment will greatly degrade the speech recognition performance. Thus, how to find the solutions for reducing the environmental mismatch is a very important topic. The codeword-dependent cepstrum normalization (CDCN) [1] was a solution using the maximum likelihood (ML) criterion and can effectively compensate for the environmental mismatch. The RelAtive SpecTrAl (RASTA) [2] was also proved to be well-performed for a recognizer trained by the clean speech and tested by data with both convolutional and additive noises. Cepstral mean normalization (CMN) [3] is another efficient method for telephone speech recognition because it can suppress the slow-varying channel effect of telephone speech. This method was also validated for a speaker verification system over telephone networks [4]. Sankar and Lee [5] proposed a maximum likelihood stochastic matching method to reduce the acoustic mismatch between a testing utterance and a given reference model. Rahim and Juang [6] proposed a signal bias removal method for removing the environmental bias by iteratively estimating the bias between training model and newly bias-removed testing utterance. A rapid environment adaptation algorithm based on spectrum equalization [7] was proposed to extract the spectra of noises between testing and training environments.

In this paper, we focus on solving the problem of channel effect in telephone speech. We propose a channel estimation method based on the maximum a posteriori (MAP) criterion. It is to maximize a posteriori probability of the channel cepstral vector conditioned on a given state sequence. The derived channel cepstral vector includes an interpolation factor of a priori channel statistics. A priori channel statistics are determined by 70 estimated channel cepstral vectors, which are extracted from telephone speech. In this study, the reference models are trained by using clean speech database. When a telephone speech is matched to the reference models, the channel cepstral vectors are estimated. Then, the reference models are adapted by adding the estimated channel cepstral vectors. Finally, the telephone speech is recognized using the adapted reference models.

2. CHANNEL ESTIMATION FOR REFERENCE MODEL ADAPTATION

2.1 System Overview

The proposed system structure is shown in Figure 1. All feature variables are represented in cepstral domain. There are two stages of Viterbi decoding in the figure. The telephone speech cepstral vector $Y = \{y_1, y_2, \cdots, y_T\}^t$, where t and T are transpose operator and frame number respectively, is time aligned by the first pass Viterbi decoder to obtain the state sequence $S^m = \{s_1^m, s_2^m, \cdots, s_T^m\}^t$ for m^{th} reference model. However, in order to align a better state sequence, the Viterbi decoder is merged with a priori channel Gaussian parameters (μ_{ch}, Σ_{ch}). Let $S = \{S^1, S^2, \cdots, S^M\}$ be the decoded state matrix for M reference models. By using the MAP estimation theory, the channel cepstral matrix $V_{ch} = \{v_{ch}^1, v_{ch}^2, \cdots, v_{ch}^M\}$ is then determined.

During the estimation process, a priori channel parameters (μ_{ch}, Σ_{ch}) are also included. The clean reference models $\Lambda_x = \{\Lambda_x^1, \Lambda_x^2, \cdots, \Lambda_x^M\}$ are then added by the estimated channel cepstral matrix V_{ch} to generate an updated reference models $\Lambda_y = \{\Lambda_y^1, \Lambda_y^2, \cdots, \Lambda_y^M\}$, i.e. $\Lambda_y = \Lambda_x + V_{ch}$. Finally, the telephone speech is recognized by using the adapted reference models.

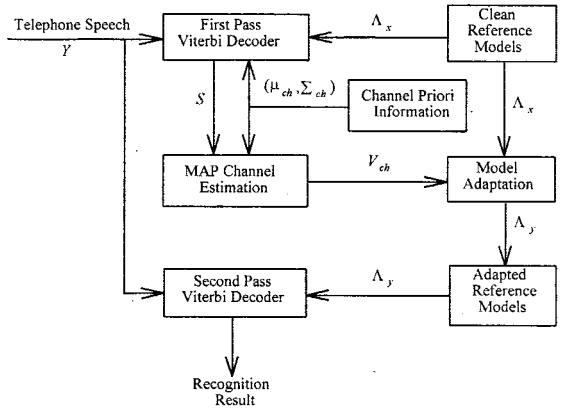

Figure 1 System structure for telephone speech recognition

2.2 Formula Derivation

In this study, maximum a posteriori estimation is applied for channel estimation. Let the variable v_{ch}^m denote the channel cepstral vector of a telephone utterance Y matched to m^{th} reference model. The channel cepstral vector is assumed to be Gaussian distributed with mean vector μ_{ch} and covariance matrix Σ_{ch}. A posteriori probability of channel cepstral vector given a state sequence $S^m = \{s_1^m, s_2^m, \cdots, s_T^m\}^t$ is expressed as $P(v_{ch}^m | Y, S^m)$, where the cepstral vector of a hidden Markov model (HMM) state s_t^m is also modeled as the Gaussian distribution with mean vector $\mu_{s_t^m}$ and covariance matrix $\Sigma_{s_t^m}$. Then, the MAP estimation of channel cepstral vector $\hat{v}_{ch}^m$ can be obtained by maximizing a posteriori probability $P(v_{ch}^m | Y, S^m)$. That is,

$$\hat{v}_{ch}^m = \underset{v_{ch}^m}{\operatorname{argmax}} P(v_{ch}^m | Y, S^m) \tag{1}$$

Applying Bayes' rule, Eq.(1) is expressed as

$$\begin{aligned} \hat{v}_{ch}^m &= \underset{v_{ch}^m}{\operatorname{argmax}}\ k \cdot P(Y | v_{ch}^m, S^m) \cdot P(v_{ch}^m | S^m) \\ &= \underset{v_{ch}^m}{\operatorname{argmax}}\ k \cdot P(Y | v_{ch}^m, S^m) \cdot P(v_{ch}^m) \end{aligned} \tag{2}$$

where $k = 1/P(Y|S^m)$ and state sequence S^m is assumed to be independent of v_{ch}^m. If a priori channel probability $P(v_{ch}^m)$ is excluded from the equation, the resulting estimation criterion becomes the maximum likelihood criterion. In this study, a priori channel probability $P(v_{ch}^m)$ will be optionally included in the estimation of channel cepstral vector for comparative study. To derive the channel cepstral vector $\hat{v}_{ch}^m$, we discard the constant k in Eq.(2) and maximize the logarithm of the following probability

$$\begin{aligned} &\ln\{P(Y | v_{ch}^m, S^m) \cdot P(v_{ch}^m)\} = \\ &\ln\{P(y_1, y_2, \ldots, y_T | v_{ch}^m, s_1^m, s_2^m, \cdots, s_T^m) \cdot P(v_{ch}^m)\} \end{aligned} \tag{3}$$

If the observation Y is independent for each frame, the joint probability function is the product of the probabilities of each frame. Eq.(3) is rewritten as

$$\ln\{P(Y | v_{ch}^m, S^m) \cdot P(v_{ch}^m)\} = \ln\left(\prod_{t=1}^{T} \{P(y_t | v_{ch}^m, s_t^m) \cdot P(v_{ch}^m)\}\right) \tag{4}$$

When the probabilities are expressed in Gaussian density, the right side of Eq.(4) is expressed as

$$\begin{aligned} \sum_{t=1}^{T} \{ & k_{s_t^m} - \frac{1}{2}(y_t - \mu_{s_t^m} - v_{ch}^m)^t \Sigma_{s_t^m}^{-1} (y_t - \mu_{s_t^m} - v_{ch}^m) \\ & + k_{ch} - \frac{1}{2}(v_{ch}^m - \mu_{ch})^t \Sigma_{ch}^{-1} (v_{ch}^m - \mu_{ch}) \} \end{aligned} \tag{5}$$

where $k_{s_t^m}$ and k_{ch} are normalization constants. Then, the MAP estimation of channel cepstral vector $\hat{v}_{ch}^m$ is obtained by finding the gradient of Eq.(5) with respect to channel cepstral vector v_{ch}^m and setting the result to zero. The MAP estimation of channel cepstral vector $\hat{v}_{ch}^m$ is derived as

$$\hat{v}_{ch}^m = \left(\sum_{t=1}^{T} \Sigma_{s_t^m}^{-1} + T \cdot \Sigma_{ch}^{-1}\right)^{-1} \cdot \left(\sum_{t=1}^{T} \Sigma_{s_t^m}^{-1} (y_t - \mu_{s_t^m}) + T \cdot \Sigma_{ch}^{-1} \cdot \mu_{ch}\right) \tag{6}$$

From the above equation, we know that the channel effect in telephone speech is estimated according to the cepstral difference between testing utterance and decoded state sequence. Thus, we apply the channel a priori information in the first pass Viterbi decoding for a correct time alignment to estimate an accurate channel cepstral vector $\hat{v}_{ch}^{m}$. In addition, it is reasonable that the cepstral difference is weighted by the state covariance matrix. We can find that the channel estimator is interpolated by a priori channel statistics which can increase the estimation accuracy. If a priori channel information is not included in the MAP channel estimation, it becomes the ML estimation.

$$\tilde{v}_{ch}^{m} = \underset{v_{ch}^{m}}{\operatorname{argmax}}\ P(Y|v_{ch}^{m}, S^{m}) = \left(\sum_{t=1}^{T}\Sigma_{s_t^m}^{-1}\right)^{-1}\cdot\left(\sum_{t=1}^{T}\Sigma_{s_t^m}^{-1}(y_t-\mu_{s_t^m})\right) \tag{7}$$

2.3 A Priori Channel Statistics

In this study, a priori channel statistics plays an important role in the MAP channel estimation. We propose an iterative procedure for estimating a priori channel statistics. First, we extract 70 telephone utterances to estimate the channel cepstral vectors and determine a priori channel mean vector μ_{ch} and covariance matrix Σ_{ch}. For each telephone utterance, we apply the following iterative procedure:

(1) The correct reference states refering to the telephone utterance are extracted. Set iteration index $i=1$.

(2) The state sequence $\{s_1^{(i)}, s_2^{(i)}, \cdots, s_T^{(i)}\}'$ is found by applying the Viterbi decoding for the telephone utterance and correct reference state models.

(3) Compute the maximum likelihood of channel cepstral vector $v_{ch}^{(i)}$ for i^{th} iteration by the following equation

$$\tilde{v}_{ch}^{(i)} = \left(\sum_{t=1}^{T}\Sigma_{s_t^{(i)}}^{-1}\right)^{-1}\cdot\left(\sum_{t=1}^{T}\Sigma_{s_t^{(i)}}^{-1}(y_t-\mu_{s_t^{(i)}})\right) \tag{8}$$

where $\mu_{s_t^{(i)}}$ and $\Sigma_{s_t^{(i)}}$ are state mean vector and covariance matrix for i^{th} decoded state sequence.

(4) The reference model is updated by adding the estimated channel cepstral vector ($\mu_{s_t^{(i)}} \rightarrow \mu_{s_t^{(i)}} + \tilde{v}_{ch}^{(i)}$). Replace i by $i+1$ and go to step 2 for next iteration.

(5) The procedure stops when the estimated vector converges or the preset iteration number is met.

Then, a priori channel parameters are determined by the estimated channel cepstral vectors. Figure 2 shows several samples of estimated channel spectra. We can see that the estimated spectra are close to the characteristics of telephone channel as we expect.

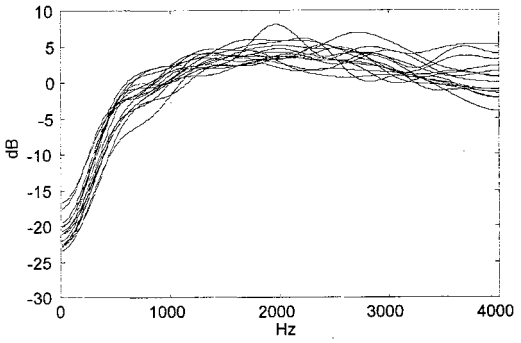

Figure 2 Several samples of estimated channel spectra

3. EXPERIMENTAL RESULTS

There are two databases used in the experiments. One is for training and the other is for testing. The speech signal is sampled at 8 kHz. For each speech frame, the feature vector consists of 12-order cepstral coefficients, 12-order delta cepstral coefficients, 1 delta log energy and 1 delta delta log energy. In this study, only 12-order cepstral coefficients are employed in reference model adaptation. The speech recognizer is based on the conventional continuous-density HMM. Each HMM's state is modeled as one mixture Gaussian density. In the training database, there are 5045 Mandarin words spoken by 51 males and 50 females. The training data is recorded by a high-quality microphone for generating the clean reference models. The model parameters are trained by using the segmental k-means algorithm. We use subword models as reference models. In testing phase, a multispeaker (35 males and 35 females) speech recognition task for 1075-Mandarin-name is conducted to evaluate the proposed method. The testing database is composed of 3006 utterances which are recorded through the telephone networks by using 10 telephone handsets. The channel a priori parameters μ_{ch} and Σ_{ch} are determined by using 70 telephone utterances, which are not appeared in the testing database. The result of cepstral mean normalization method [3] is included for comparison. The proposed reference

model adaptation with and without channel a priori information are also presented in the experiments.

The top n recognition rates are listed in Table 1. We can see that the recognition performance is poor when the training and testing environments are mismatch. The cepstral mean normalization method can greatly increase the top 5 recognition rates from 49.4% to 85.8%. This shows the effectiveness of cepstral mean normalization. When the proposed adaptation method without a priori channel information is applied, i.e. by using ML criterion for channel estimation, the recognition performance is slightly worse than that of cepstral mean normalization. But, when the model adaptation method with a priori channel information is applied, i.e. by using MAP criterion for channel estimation, the performance is better than that of cepstral mean normalization. The MAP channel estimation outperforms the ML channel estimation because the MAP channel estimation applies more information for channel estimation.

. Table 1 Comparison of top n recognition rates

	Top 1	Top 2	Top 3	Top 4	Top 5
No Adaptation	25.0	34.3	40.2	44.9	49.4
CMN	63.2	74.3	79.9	83.6	85.8
Chan. Est. by ML	61.1	72.9	78.7	82.1	84.7
Chan. Est. by MAP	63.6	74.9	80.8	84.8	87.1

4. CONCLUSION

This paper presents a channel estimation method for reference model adaptation in telephone speech recognition. We derive a MAP channel estimation formula for adapting reference models. The channel cepstral vector is estimated by maximum a posteriori probability of channel cepstral vector given the state sequence. The resulting estimated channel cepstral vector includes an interpolation factor of a priori channel statistics, which can compensate the estimation error. Thus, for a telephone utterance, the channel cepstral vector is estimated by the first Viterbi decoder. The reference models are then adpated by the estimated channel vectors. The recognition results are obtained by using the adapted reference models. Experiments show that a priori channel statistics plays an important role in the channel estimation and the proposed method can significantly improve the recognition rates for telephone speech recognition.

ACKNOWLEDGEMENT

This research has been partially supported by National Science Council, Taiwan, R.O.C. under contract no. NSC-84-0404-E-007-056.

REFERENCES

[1] A. Acero and R. M. Stern, "Environmental Robustness in Automatic Speech Recognition", Proc. ICASSP, 1990, pp. 849-852.

[2] H. Hermansky, N. Morgan and H. G. Hirsch, "Recognition of Speech in Additive and Convolutional Noise Based on RASTA Spectral Processing", Proc. ICASSP, 1993, Vol 2, pp. 83-86.

[3] C. Mokbel, P. Paches-leal, D. Jouvet and J. Monne, "Compensation of Telephone Line Effect for Robust Speech Recognition", Proc. ICSLP, 1994, pp. 987-990.

[4] A. E. Rosenberg, C. H. Lee and F. K. Soong, "Cepstral Channel Normalization Techniques for HMM-Based Speaker Verification", Proc. ICSLP, 1994, pp. 1835-1838.

[5] A. Sankar and C. H. Lee, "Robust Speech Recognition Based on Stochastic Matching", Proc. ICASSP, 1995, pp. 121-124.

[6] M. G. Rahim and B. H. Juang, "Signal Bias Removal for Robust Telephone Based Speech Recognition in Adverse Environments", Proc. ICASSP, 1994, Vol 1, pp. 445-448.

[7] K. Takagi, H. Hattori and T. Watanabe, "Speech Recognition with Rapid Environment Adaptation by Spectrum Equalization", Proc. ICSLP, 1994, pp. 1023-1026.