ADAPTATION METHOD BASED ON HMM COMPOSITION AND EM ALGORITHM

Yasuhiro Minami and Sadaoki Furui
NTT Human Interface Laboratories
Musashino-shi, Tokyo, 180 Japan

ABSTRACT

A method for adapting HMMs to additive noise and multiplicative distortion at the same time is proposed. This method first creates a noise HMM for additive noise, then composes HMMs for noisy and distorted speech data from this HMM and speech HMMs so that these composed HMMs become the functions of signal-to-noise (S/N) ratio and multiplicative distortion. S/N ratio and multiplicative distortion are estimated by maximizing the likelihood of the HMMs to the input speech. To achieve this, we propose a new method that divides the maximization process into estimation of S/N ratio and estimation of cepstrum bias. The S/N ratio is estimated using the parallel model method. The cepstrum bias is estimated using the EM algorithm. To evaluate this method, two experiments in terms of phoneme recognition and connected digit recognition are performed. The guarantee of convergence of this algorithm is also discussed.

1. INTRODUCTION

Background noise, channel noise, and channel distortion are serious problems in speech recognition. They can generally be modeled by combining additive noise and multiplicative distortion in the linear spectral domain. If adaptation for both additive noise and multiplicative distortion can be performed simultaneously in the speech model, effective speech recognition can be achieved.

Many studies have been done on adaptation for either multiplicative distortion or additive noise [1]-[6]. For multiplicative distortion, CMN (cepstral mean normalization), which regards the average of speech cepstra as multiplicative distortion, has been proposed [1]. Since CMN is a simple and powerful adaptation technique, it is widely used in many recognition systems. Sankar proposed the cepstrum bias (multiplicative distortion) estimation method based on the Maximum Likelihood (ML) method [2]. This method was applied to telephone speech and proved effective for multiplicative distortion. For additive noise, HMM decomposition, parallel model combination (PMC), and HMM composition were proposed [4][5][6]. These methods approximate the distributions of the random variables for noisy speech from the distributions of the random variables for speech and noise.

These techniques treat either multiplicative distortion or additive noise. It is difficult to achieve adaptation for additive noise and multiplicative distortion at the same time, because nonlinear transformation occurs between cepstral coefficients and the linear spectrum. To solve this problem, we previously proposed a method that estimates additive noise and multiplicative distortion by maximizing the likelihood of HMMs [7]. However, our previous method required complex calculations to do this.

This paper describes a method that can maximize the likelihood more easily. Modeling of noisy and distorted speech in the linear power spectrum domain is discussed first. Then an adaptation method for additive noise and multiplicative distortion is discussed.

2. NOISY AND DISTORTED SPEECH MODELING

Our model for speech signals in general noisy conditions

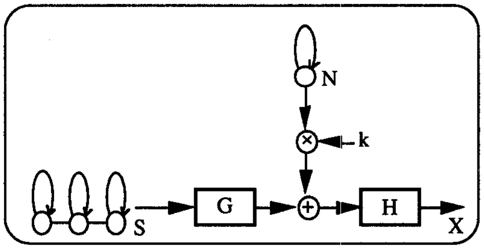

Figure 1. Model for producing noisy and distorted speech.

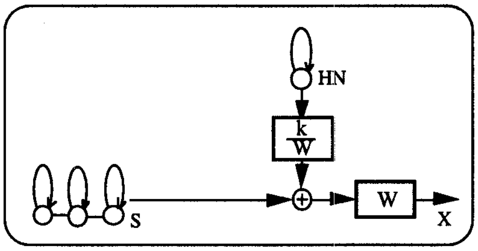

Figure 2. Converted model for producing noisy and distorted speech.

0-7803-3192-3/96 $5.00©1996 IEEE

is shown in Figure 1. Speech signal ***S*** is produced by speech HMMs. Noise signal ***N*** is produced by a noise HMM. ***S*** and ***N*** are defined in the linear power spectral domain. ***S*** is first multiplied by multiplicative distortion ***G*** corresponding to a change in speaking style, etc. Then additive noise ***N*** is added to speech signal ***GS***. At this time, ***N*** is multiplied by coefficient *k* corresponding to the power ratio between speech and noise. Finally, the speech signal is multiplied by multiplicative distortion ***H*** corresponding to line distortion, microphone distortion, etc. Using these notations, the final noisy and distorted speech signal is $X = H(GS + kN) = HGS + kHN$. By setting $W = HG$, $X = WS + kHN = W(S + kHN/W)$ is obtained, so the basic noisy speech model can be converted into the model shown in Figure 2. The HMM for ***HN*** can be trained by using a signal without speech. The HMMs for ***S*** can be made from noise-free data. If *k* and ***W*** can be estimated, HMMs that generate ***X*** can be obtained. The problem is how to estimate *k* and ***W***.

3. ADAPTATION FORMULATION

Ordinary HMM composition considered only additive noise. To estimate the values of *k* and ***W***, we extended ordinary HMM composition so that composed HMMs may become a function of *k* and ***W***. A set of phoneme HMMs producing ***X*** is modeled by composing the ***kHN*** / ***W*** HMM and the ***S*** HMMs by the extended HMM composition. ***W*** and *k* are then estimated by maximizing the trellis likelihood score $P(O|M(k,W))$, where $O = \{x_1, x_2, ..., x_T\}$ is a time sequence of input vectors and $M(k,W)$ is a set of composed phoneme models as functions of *k* and ***W***. In our previous method, the steepest descent method was used to maximize $P(O|M(k,W))$. This method estimated *k* and ***W*** simultaneously, but it needed complex equations. In its place we propose a new method that divides the maximization process into *k* estimation and ***W*** estimation.

The *k* is estimated using the parallel model method. In this method, several sets of models with different k_j's are prepared. Using these models, the likelihood scores, $P(O|M(k_j,W))$, are calculated for all *j*'s, and a set of models with maximum likelihood is selected. The ***W*** is estimated using the EM algorithm.

Thus, our algorithm to find both *k* and ***W*** is as follows:

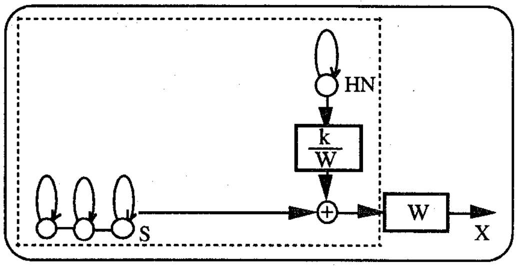

Figure 3. Implementation of proposed method.

1. Initialize ***W***.
2. Compose sets of HMMs, changing *k*; select the *k* that gives the maximum value of $P(O|M(k,W))$.
3. Compose the set of HMMs in the area bounded by the dotted line in Figure 3 with the fixed *k*/***W*** obtained in step 2.
4. Estimate ***W*** outside the dotted line by using Sankar's cepstral bias estimation method.
5. Update *k*/***W*** inside the dotted line using the newly estimated ***W***.
6. Repeat steps 2 to 5 until convergence is achieved.

Sankar's cepstral bias estimation method estimates bias ***b*** between x_t and y_t by using the EM algorithm, where x_t is the input vector and y_t is the corresponding vector produced from speaker independent HMMs; ***b*** is the inverse cosine transformation of the log transformation of ***W***. The equation for estimating ***b*** is

$$b_i = \frac{\sum\limits_{t,n,m,l}^{T,N,M,L} \gamma_t(n,m,l)(x_{t,i} - \mu_{n,m,l,i}) / \sigma^2_{n,m,l,i}}{\sum\limits_{t,n,m,l}^{T,N,M,L} \gamma_t(n,m,l) / \sigma^2_{n,m,l,i}}, \quad (1)$$

where *t* is time, *n* and *m* are state numbers, *l* indicates the *l*th mixture component, and *i* indicates the *i*th element of the vector; $\gamma_t(n, m, l)$ is the joint probability of observing ***O*** and transiting from state *n* to state *m* from time *t* to time *t*+1 and taking the *l*th mixture component to produce x_t. $\gamma_t(n, m, l)$ can be calculated using the forward-backward algorithm.

4. EXPERIMENTS

Two experiments were performed to evaluate our method. One compared our method with ordinary HMM composition, our previous method, and Sankar's method in terms of phoneme recognition. The other compared our method with CMN (cepstrum mean normalization) and ordinary HMM composition in terms of connected digit recognition.

(1) Phoneme recognition experiment

The speaker-independent HMMs were trained using speech data uttered by 64 speakers under noise-free conditions. The HMMs had three states, each with a four-mixture Gaussian distribution. The noise HMM had one state and a single Gaussian distribution. One sentence, with the transcription, uttered by one male speaker was used for adaptation. This means that adaptation was performed in a supervised mode. The evaluated sentences were 51 telephone directory inquiry sentences and were uttered by the same male speaker. Noise recorded in a computer room was added to each speech data at 6 and 12 dB signal-to-noise (S/N) ratio. The data was then passed through a filter whose characteristic was $1\text{-}0.97z^{-1}$. The input data was sampled at 12 kHz. Although input features consisting of 16-order cepstrum, 16-order delta cepstrum, and one delta power were used, only the cepstrum distributions were adapted to the noisy and distorted speech. We compared five methods:

(a) No adaptation.
(b) Ordinary HMM composition: k was set so that it gave the maximum likelihood to the adaptation speech.
(c) Sankar's method: The cepstrum bias (multiplicative distortion) was estimated by directly applying Sankar's method to the adaptation speech.
(d) Our method using the steepest descent method: This method was proposed in our previous paper [7]. k and $\boldsymbol{W}$ were estimated by maximizing the likelihood using the steepest descent method.
(e) Our method using the EM algorithm: k and $\boldsymbol{W}$ were estimated by our new method.

Figure 4 shows the experimental results. Methods (b) and (c) consider only additive noise or multiplicative distortion. Therefore, while both methods showed improvement in the recognition rate, the degree of improvement was small. On the other hand, our method showed great improvement for both S/N ratios. Our method showed almost the same performance using the EM algorithm as using the steepest descent method, and the only complicated equation required in our new method is Eq. (1).

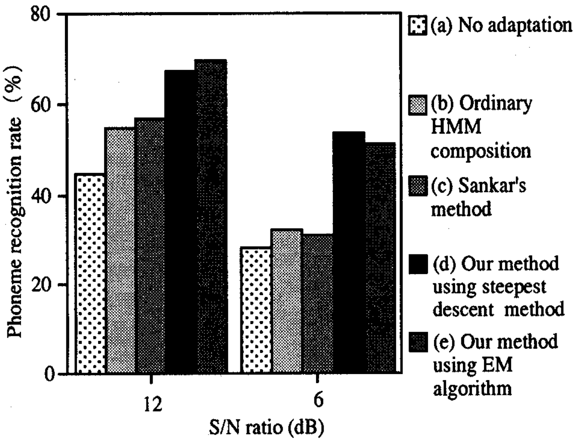

Figure 4. Phoneme recognition rate for each adaptation method.

(2) Connected digit recognition experiment

In this experiment, four-digit strings were recognized when the number of digits was known. 9702 digit strings uttered by 70 male speakers were used for training speaker-independent HMMs. A whole-word HMM was prepared for each digit. The number of states depends on the digits. The number of mixture components in each state was set to four. 16-order cepstrum and 16-order delta cepstrum were used. The delta power was not used here.

Adaptation was performed in an unsupervised mode. First a universal speech HMM was made by using all the training speech data. The speech HMM had one state with 16-mixture Gaussian distribution. S/N ratio and multiplicative distortion were then estimated using the proposed algorithm from the speech HMM and noise HMM. Finally, HMMs with additive noise and multiplicative distortion were created from the noise HMM and the digit HMMs using the estimated S/N ratio and multiplicative distortion.

1750 digit strings uttered by 50 speakers were simultaneously recorded using two microphones, a condenser microphone and a boundary microphone. The characteristics of the boundary microphone were very different from those of the microphone used in training.

Two types of noise, computer room noise and car noise, were recorded using the same two microphones in an anechoic room. In the experiment, speech data were made by adding these to clean speech data so that the S/N ratios became 12 and 18 dB. In the experiment, CMN and ordinary HMM composition were compared with our method. Since the weight value when the noise data added to the speech data was made constant for the entire speech data over all speakers, the S/N of each speaker varied according to the loudness of voice (at 12 dB, it varied from 4 to 17 dB). In ordinary HMM composition, k was estimated at each utterance by using step 2 in section 3.

Table 1 shows the base-line recognition rates for the clean speech data. The recognition rate dropped by about 4% using the boundary microphone, compared to that with the condenser microphone. This was mainly due to the frequency characteristics of the boundary microphone.

Table 2 shows the result for computer room noise using the boundary microphone. The recognition rate of our method was better, by 1.1 % at 12 dB and by 2.1 % at 18 dB, than the ordinary HMM composition method. Moreover, it shows a great improvement over CMN.

Table 1. Base-line digit string recognition rate for clean speech.

Condenser microphone	Boundary microphone
97.3%	93.0%

Table 3 shows the results for car noise using the boundary microphone. In this case, our method also improved the recognition rate by 1.8 % at 12 dB and 4.2 % at 18 dB.

5. DISCUSSION

When an iterative algorithm is proposed, it should express a guarantee to converge the values. Unfortunately, our algorithm does not provide such a guarantee. Here we discuss where the guarantee is not satisfied. To do this, we must check steps 2, 4, and 5 in section 3 (the other steps are not related to the guarantee of convergence). After step 2, it is obvious that the likelihood $P(O|M(k,W))$ is always bigger than before. This means that step 2 holds the guarantee. Because step 4 is the EM algorithm, after this step, $P(O|M(k,W))$ is also always bigger than before. However, there is no guarantee that $P(O|M(k,W))$ is always bigger than before when k/W is replaced using the newly estimated W in step 5. However, in our experiments, we did not see the phenomenon of $P(O|M(k,W))$ causing the fluctuation. Moreover, $P(O|M(k,W))$ never decreased significantly after step 5. This confirms that our algorithm has no problems in practical use.

6. CONCLUSION

This paper described a method that can adapt HMMs to additive noise and multiplicative distortion at the same time. In this method, HMM composition is extended so that composed HMMs may become the functions of S/N ratio and multiplicative distortion. The method creates HMMs from speech HMMs and a noise HMM by extended HMM composition. It then estimates S/N ratio and multiplicative distortion by maximizing its likelihood for input speech. This algorithm was evaluated by the phoneme recognition experiment and the connected digit recognition experiment. The phoneme recognition result showed that our algorithm can achieve the same recognition rate as our previous algorithm, which required highly complex calculation. The digit recognition results showed that our new algorithm was more effective for additive noise and multiplicative distortion than CMN and ordinary HMM composition which considers only additive noise.

Table 2. Digit string recognition rate with computer room noise (boundary microphone).

Method	12 dB	18 dB
No adaptation	25.6 %	56.9 %
CMN	33.6 %	69.9 %
Ordinary HMM composition	59.2%	82.5 %
Proposed method	60.3 %	84.6 %

Table 3. Digit string recognition rate with car noise (boundary microphone).

Method	12 dB	18 dB
No adaptation	68.4%	80.8%
CMN	82.6%	94.0%
Ordinary HMM composition	87.2%	91.8%
Proposed method	89.0%	96.0%

ACKNOWLEDGMENTS

We thank the members of the Furui Research Laboratory of the NTT Human Interface Laboratories for their useful discussions. We also thank Mr. Takagi for collecting digit string data.

REFERENCES

[1] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", *Journal of the Acoustic Society of America*, vol. 55, June 1974, pp. 1304-1312.

[2] A. Sankar and C.-H. Lee, "Robust speech recognition based on stochastic matching", *Proc. Int. Conf. Acoust. Speech, Signal Process.*, May 1995, pp. 121-124.

[3] M. G. Rahim and B.-H. Juang, "Signal bias removal for robust telephone-based speech recognition in adverse environments", *Proc. Int. Conf. Acoust. Speech, Signal Process.*, April 1994, pp. 445-448.

[4] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise", *Proc. Int. Conf. Acoust. Speech Signal Process.*, 1990, pp. 845-848.

[5] M. J. F. Gales and S. J. Young, "An improved approach to the hidden Markov model decomposition of speech and noise", *Proc. Int. Conf. Acoust. Speech Signal Process.*, March 1992, pp. 233-236.

[6] F. Martin, K. Shikano, and Y. Minami, "Recognition of noisy speech by composition of hidden Markov models", *Proc. Eurospeech*, Berlin, September 1993, pp. 1031-1034.

[7] Y. Minami and S. Furui, "A maximum likelihood procedure for a universal adaptation method based on HMM composition", *Proc. Int. Conf. Acoust. Speech, Signal Process.*, May 1995, pp. 129-132.