A maximum likelihood procedure for a universal adaptation method based on HMM composition

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

ABSTRACT

This paper proposes an adaptation method for universal noise (additive noise and multiplicative distortion) based on the HMM composition (compensation) technique. Although the original HMM composition can be applied only to additive noise, our new method can estimate multiplicative distortion by maximizing the likelihood value. Signal-to-noise ratio is automatically estimated as part of the estimation of multiplicative distortion. Phoneme recognition experiments show that this method improves recognition accuracy for noisy and distorted speech.

1. INTRODUCTION

Background noise, channel noise, and channel distortion are crucial problems in speech recognition. They are usually modeled by combining additive noise and multiplicative distortion in the linear spectral domain. Speaker characteristics can also be regarded as multiplicative distortion [1]. If both additive noise and multiplicative noise can be simultaneously estimated, that is, if universal noise adaptation can be achieved, it should to be very useful in speech recognition applications.

Various methods for removing estimated noise and distortion have been proposed, including spectral subtraction for additive noise and cepstral normalization for multiplicative distortion. However, since these methods utilize the average values of linear spectra and cepstra as noise and distortion, they cannot be simply extended to remove combinations of additive noise and multiplicative distortion.

An accurate noise adaptation method using a noise model, called HMM composition or compensation, has recently been proposed for additive noise [2][3]. This method creates HMMs for noisy conditions using speech HMMs and a noise HMM. It uses the mean and covariance of the noise distribution to adapt the speech distribution. However, it does not consider multiplicative distortion.

A different approach has been proposed to estimate either the multiplicative noise or additive noise spectrum by maximizing the likelihood value. Rahim et al. removed telephone line bias in the cepstrum domain (multiplicative distortion) [4]. Rose et al. formulated the maximum likelihood parameter estimation procedure for additive noise or multiplicative noise [5].

This paper extends the HMM composition method to accommodate both additive and multiplicative noise by using a maximum likelihood estimation criterion. In this framework, the S/N estimation is performed as part of the estimation of multiplicative distortion.

2. NOISY AND DISTORTED SPEECH MODELING

The model for producing speech signals under most noisy conditions is shown in Figure 1. Speech signal S is produced by speech HMMs and noise signal N is produced by a noise HMM. Both S and N are defined in the linear-power spectral domain. First, S is multiplied by multiplicative distortion G,

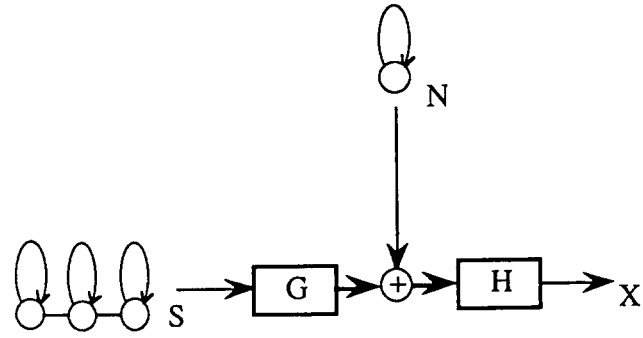

Figure 1. A model for processing noisy and distorted speech.

0-7803-2431-5/95 $4.00

which includes speaker characteristics. Then additive noise N is added to speech signal SG. Finally, the speech signal is multiplied by multiplicative noise H, which includes line and channel distortion. We thus obtain the final noisy and distorted speech signal as X (= H(GS + N) = HGS + HN). By setting W = HG, we get X = WS + HN; therefore, the basic noisy speech model can be converted into the model shown in Figure 2.

The HMM for HN can be trained by using the signal recorded for a period without speech. The HMMs for S can be made from noise-free speech. The problem is how to estimate W. Since W is multiplicative distortion, it can be written as

$$W = \{ w_0, w_1, w_2, ..., w_p \}, \quad (1)$$

in the linear spectrum domain, where p+1 is the number of power spectral components.

3. FORMULATION OF ADAPTATION

To estimate the value of W, we model X by combining the HMMs for HN and WS using the HMM composition method. Here we assume that W is a fixed vector. Then W is estimated by maximizing the likelihood score P(O|M(W)) or P(O,Λ|M(W)), where P(O|M(W)) is the trellis likelihood score, P(O,Λ|M(W)) is the Viterbi likelihood score, $O = \{x_1, x_2, ..., x_T\}$ is a time sequence of input vectors, M(W) is a set of phoneme models as functions of W, and $\Lambda = \{s_1, s_2, ..., s_T\}$ is the time sequence of states.

To maximize P(O|M(W)) or P(O,Λ|M(W)), we propose the following two methods.

(1) Exact method

The P(O,Λ|M(W)) can be maximized by using the steepest descent method, using the following iterative equation.

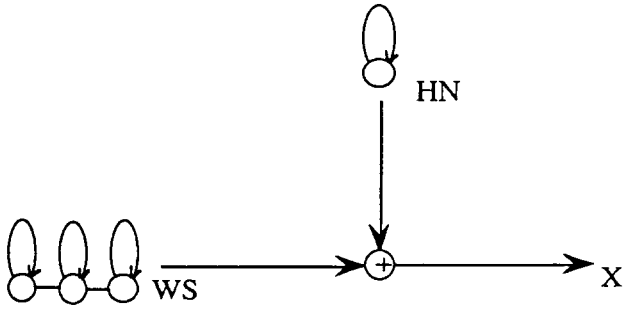

Figure 2. A converted model for processing noisy and distorted speech (W=HG).

$$\mathbf{W}(k)=\mathbf{W}(k-1)+\varepsilon\frac{\partial \log(P(O,\Lambda|M(W))}{\partial \mathbf{W}}, \quad (2)$$

where ε is the step size.

If the output probabilities are represented by mixture Gaussian densities, it becomes complicated to maximize Eq. (2) directly. Therefore the maximum single Gaussian density in each mixture is used instead of multiple mixtures.

$$b_{S_iS_j}(x_t) = \frac{1}{(2\pi)^p |\Sigma_{S_iS_j}|} e^{-\frac{1}{2}(x_t-\mu_{S_iS_j})^t \Sigma_{S_iS_j}^{-1}(x_t-\mu_{S_iS_j})} \quad (3)$$

Viterbi decoding is used to obtain P(O,Λ|M(W)), $\mu = \{\mu_1, \mu_2, ..., \mu_T\}$, and $\Sigma = \{\Sigma_1, \Sigma_2, ..., \Sigma_T\}$, where μ and Σ are the sequences of mean vectors and covariance matrices selected by the Viterbi algorithm.

To deal with many random variables in the equations, the following conventions are used. R_c represents source R in domain c, where c ={cep, lg, lin}. For instance, X_{lin} is the random variable associated with noisy speech in the linear spectrum. The corresponding Gaussian distribution is $N(\mu^{X_{lin}}, \Sigma^{X_{lin}})$. The main notations for random variables are as follows.

N_{cep}, N_{lg}, N_{lin}: Noise in the cepstrum, the logarithm spectrum, and the linear spectrum domain.

S_{cep}, S_{lg}, S_{lin}: Speech in the cepstrum, the logarithm spectrum, and the linear spectrum domain.

X_{cep}, X_{lg}, X_{lin}: Noisy and distorted speech in the cepstrum, the logarithm spectrum, and the linear spectrum domain.

From the definition for HMM composition (compensation), the following equations are obtained.

$$\mu^{S_{lg}} = \Gamma^{-1}\mu^{S_{cep}} \quad (4)$$

$$\Sigma^{S_{lg}} = \Gamma \Sigma^{S_{cep}} \Gamma^T \quad (5)$$

$$\mu_u^{S_{lin}} = \exp(\mu_u^{S_{lg}} + \frac{\sigma_{uu}^{S_{lg}}}{2}) \quad (6)$$

$$\sigma_{uv}^{S_{lin}} = \mu_u^{S_{lin}} \mu_v^{S_{lin}} (\exp(\sigma_{uv}^{S_{lg}}) - 1) \quad (7)$$

$$\mu_u^{X_{lg}} = \log(w_u \mu_u^{S_{lin}} + \mu_u^{N_{lin}}) - \log(\frac{w_u w_u \sigma_{uu}^{S_{lin}} + \sigma_{uu}^{N_{lin}}}{(w_u \mu_u^{S_{lin}} + \mu_u^{N_{lin}})^2} + 1) \quad (8)$$

$$\sigma_{uv}^{X_{lg}} = \log(\frac{w_u w_v \sigma_{uv}^{S_{lin}} + \sigma_{uv}^{N_{lin}}}{(w_u \mu_u^{S_{lin}} + \mu_u^{N_{lin}})(w_v \mu_v^{S_{lin}} + \mu_v^{N_{lin}})} + 1) \quad (9)$$

$$\mu^{X_{cep}} = \Gamma^{-1} \mu^{X_{lg}} \quad (10)$$

$$\Sigma^{X_{cep}} = \Gamma^{-1} \Sigma^{X_{lg}} \Gamma^{-1^{T}} \quad (11)$$

Γ: cosine transform, T: transpose, and u,v: parameter indices, $0 \le u,v \le p$.

Ignoring the differential coefficients in Eq. (2) calculated from the covariance matrix (considering only the differential coefficients calculated from the mean vector), we obtain the following equation, with $1 \le u \le p$:

$$\begin{aligned} w_u(k) &= w_u(k-1) \\ &+\varepsilon \frac{\partial \sum_t \{-\log((2\pi)^p |\Sigma_t^{X_{cep}}|) - \frac{1}{2}(x_t - \mu_t^{X_{cep}})^T \Sigma_t^{X_{cep}^{-1}} (x_t - \mu_t^{X_{cep}})\}}{\partial w_u} \\ &= w_u(k-1) + \\ &\varepsilon \sum_t \{\frac{1}{2}(\frac{\partial \mu_t^{X_{cep}}}{\partial w_u})^T \Sigma_t^{X_{cep}^{-1}} (x_t - \mu_t^{X_{cep}}) + \frac{1}{2}(x_t - \mu_t^{X_{cep}})^T \Sigma_t^{X_{cep}^{-1}} \frac{\partial \mu_t^{X_{cep}}}{\partial w_u}\} \\ &= w_u(k-1) + \\ &\varepsilon \sum_t \{\frac{1}{2}(\Gamma^{-1} \frac{\partial \mu_t^{X_{lg}}}{\partial w_u})^T \Sigma_t^{X_{cep}^{-1}} (x_t - \mu_t^{X_{cep}}) + \frac{1}{2}(x_t - \mu_t^{X_{cep}})^T \Sigma_t^{X_{cep}^{-1}} \Gamma^{-1} \frac{\partial \mu_t^{X_{lg}}}{\partial w_u}\} \end{aligned} \quad (12)$$

By differentiating Eq. (8), we obtain

$$\frac{\partial \mu_u^{X_{lg}}}{\partial w_u} = \frac{\mu_u^{S_{lin}}}{\mu_u^{N_{lin}} + w_u \mu_u^{S_{lin}}} - \frac{w_u \mu_u^{N_{lin}} \sigma_u^{S_{lin}} - \mu_u^{S_{lin}} \sigma_u^{N_{lin}}}{(\mu_u^{N_{lin}} + w_u \mu_u^{S_{lin}})^3 + (\mu_u^{N_{lin}} + w_u \mu_u^{S_{lin}})(\sigma_u^{N_{lin}} + w_u^2 \sigma_u^{S_{lin}})} . \quad (13)$$

Finally, Eq. (12) can be calculated by using Eq. (13) .

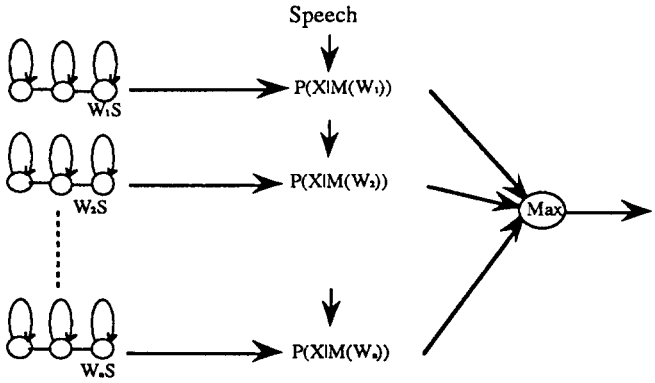

Figure 3. Parallel method.

(2) **Parallel method**

The parallel method (Figure 3) is mathematically simpler but computationally more costly than the exact. In this method, several sets of models having different W_i's are prepared. Using these models, the likelihood scores, $P(X|M(W_i))$, are calculated for all i's, and a set of models having maximum likelihood is selected.

This method is especially useful when the S/N ratio is estimated. Estimating the S/N ratio is a special case of estimating W where $W_i = \{k_1, k_1, ..., k_1\}$. Various k_i's are prepared at several intervals in the S/N ratio to estimate the S/N ratio.

4. EXPERIMENTS

We tested our method in terms of the phoneme recognition rate. Noisy speech data were artificially created on a computer, as shown in Figure 4. Noise recorded in a computer room was added to clean speech data at 12 dB. The data were then passed through a distortion filter whose characteristic was set to $1\text{-}0.97z^{-1}$; the input data were sampled at 12 kHz. The input feature vector consists of 16 cepstra, 16 delta cepstra, and 1 delta power.

A diagram of our adaptation method is shown in Figure 5. The speaker-independent HMMs were trained with speech data recorded from 64 speakers under noise-free conditions. One sentence with the transcription was used for adaptation. The training sentence was the first sentence in the phoneme-balanced sentence set. The S/N ratio was first estimated by using the parallel method. This value was then used as the initial value for the next estimation. Then, W was estimated using the exact method. Using these two steps speeds up the convergence of the likelihood value.

In the exact method, Viterbi decoding is performed first to obtain μ (= $\{\mu_1, \mu_2, ..., \mu_T\}$) and Σ (= $\{\Sigma_1, \Sigma_2, ..., \Sigma_T\}$). Then a

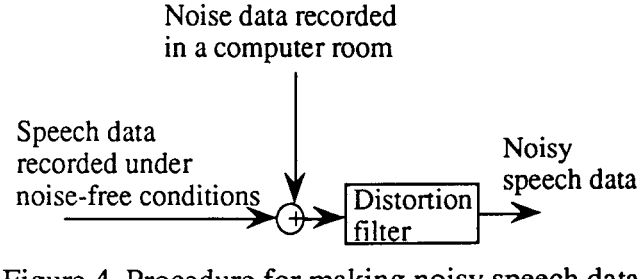

Figure 4. Procedure for making noisy speech data.

new W is calculated. These steps are repeated until the likelihood value converges, at which point the HMMs for noisy and distorted speech are obtained.

Using these HMMs, we performed phoneme recognition experiments. Fifty-one evaluation sentences were uttered by one male speaker. Since our database has only phoneme descriptions and no precise phoneme labeling, we used an evaluation algorithm for the phoneme recognition rate that does not need precise phoneme labeling [6] .

5. RESULTS

As shown in Table 1, the phoneme recognition rate for noise-added speech improved from 58.8% to 72.1% when using the S/N estimated by the parallel method. This indicates that the parallel method works well in estimating the S/N ratio. The phoneme recognition rate increased even more from 72.1% to 75.0% when we used the estimated W. This suggests that what our method estimates is a kind of speaker characteristic.

For the filtered distorted speech, the phoneme recognition rate increased from 44.7% to 56.7% when the estimated S/N ratio was used, and the recognition rate greatly increased from 56.5% to 67.7% when the estimated W was used. This means that our method can effectively estimate the filter characteristic.

6. CONCLUSION

Our proposed noise adaptation method estimates both additive noise and multiplicative distortion in a single framework. It estimates the multiplicative distortion by maximizing the likelihood value of a training sentence, based on the HMM composition technique. Phoneme-recognition experiments confirmed that this method greatly improves the recognition rate for noisy and distorted speech data.

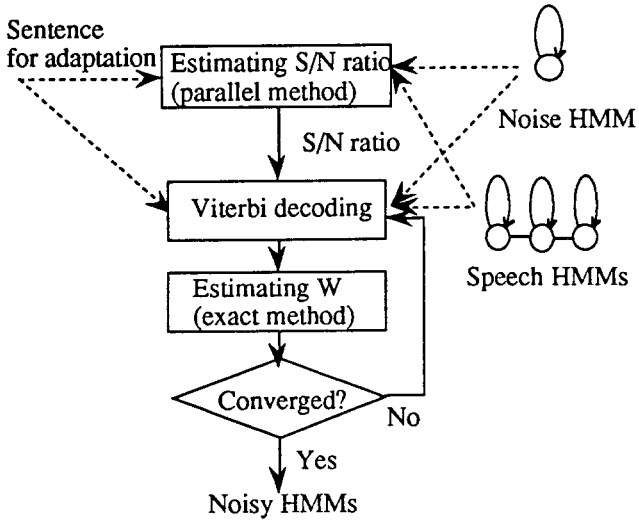

Figure 5. Procedure for making noisy HMMs.

Acknowledgments

We thank the members of the Furui Research Laboratory of the NTT Human Interface Laboratories for their useful discussions. We used the Acoustic Society of Japan's speaker-independent continuous-speech database to train the speaker-independent HMMs.

References

[1] Y. Zhao, "An acoustic-phonetic-based speaker adaptation technique for improving speaker-independent continuous speech recognition", IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 3, July 1994, pp. 380-394.

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

[3] 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.

[4] M. G. Rahim and B-H Juang, "Signal bias removal for robust telephone based speech recognition in adverse environments", Proc. ICASSP'94, April 1994, pp. I445-I448.

[5] R. C. Rose, E. M. Hofsterrer, and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise", IEEE Trans. on Speech and Audio Processing, Vol. 2, April 1994, pp. 245-257.

[6] Y. Minami, T. Matsuoka, and K. Shikano, "Phoneme evaluation algorithm without phoneme labeling", Proc. ICSLP'92, October 1992, pp. 1535-1538.

Table 1. Phoneme recognition results.

Test data	Clean HMM	HMM composition with estimated S/N ratio	HMM composition with estimated W
Noisy speech (12dB)	58.8%	72.1%	75.0%
Noise (12dB)+ multiplicative distortion	44.7%	56.5%	67.7%