IMPROVED EXTENDED HMM COMPOSITION BY INCORPORATING POWER VARIANCE

Yasuhiro Minami and Sadaoki Furui

NTT Human Interface Laboratories

Musashino-shi, Tokyo, 180 Japan

ABSTRACT

This paper describes a way of improving extended HMM composition that can precisely adapt HMMs to both noisy and distorted speech. To do this, we incorporate the variance of power into extended HMM composition using quantization to approximate the Gaussian distribution of the 0th order cepstrum. Consequently, a distribution of noisy speech is approximated in the linear spectral domain as a mixture of log normal distributions.

This method is evaluated by a four-digit recognition experiment when the number of digits is known. Two types of noise, computer room noise and car noise, are used and noisy and distorted speech data is made by adding these types of noise to speech data recorded using a boundary microphone. Results show that the proposed method improves recognition rates for noisy and distorted speech compared with our previous method.

1. INTRODUCTION

Several new noise adaptation techniques, called HMM decomposition, PMC, or HMM composition that create HMMs for noisy conditions from speech HMMs and a noise HMM have been proposed [1][2][3]. These methods approximate the distributions of the random variables for noisy speech from the distributions of the random variables for speech and noise. We showed that the HMM composition technique performed very well for continuous speech recognition.

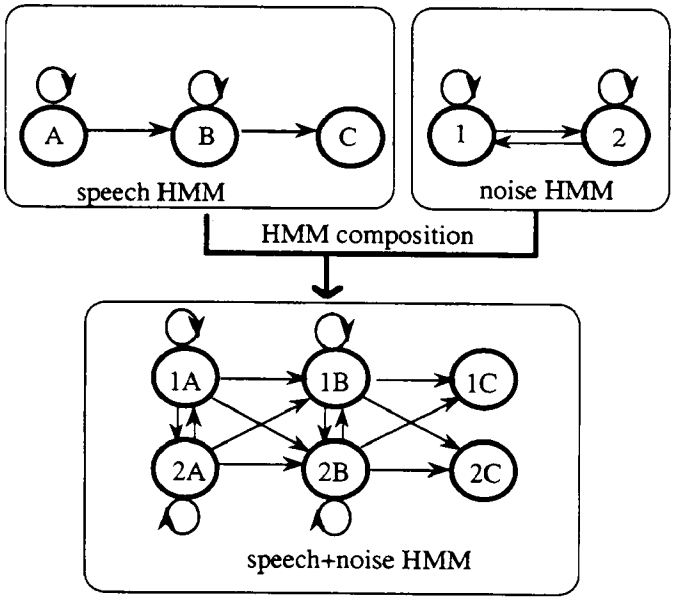

Figure 1: HMM composition.

However, these techniques treat only additive noise. It is difficult to achieve adaptation for both 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 proposed a method that estimates additive noise and multiplicative distortion by maximizing the likelihood of HMMs [4][5]. We call it extended HMM composition.

In HMM composition, so far we have not used the variance of the 0th order cepstrum as the theory describes, since using it in our experiments degraded recognition accuracy. In this paper, we describe how to incorporate power (0th order of cepstrum coefficients (cep0)) variance into HMM composition. In Section 2, we describe the HMM composition method. In Section 3, we describe extended HMM composition. In Section 4, we discuss how to incorporate cep0 variance into HMM composition.

2. HMM COMPOSITION

The HMM composition is the same as PMC proposed by Gales and Young. The basic concept is to make HMMs for noisy conditions from speech HMMs and a noise HMM (Figure 1). Speech HMMs are modeled from noise-free speech and the noise HMM is modeled from environment noise. The structure of the resulting model is a combination of the speech HMM and noise HMM. The main difficulty in combining two models is calculating the Gaussian distributions of output probabilities from the two source HMM distributions: the noise HMM and the speech HMM. Since each HMM is defined in the cepstrum domain, and speech and noise are additive in the linear spectrum domain, the Gaussian distributions defined in the cepstrum domain are transformed into log normal distributions in the linear spectrum domain and convoluted and re-transformed. This method is formulated as follows.

To deal with many random variables in the equations, the following

conventions are used. R_c represents source R in domain c, where $c = \{cep, lg, lin\}$, $R = \{S, N, X\}$. S, N and X are random variable vectors of speech, noise, and noisy speech respectively. For instance, X_{lin} is the random variable associated with noisy speech in the linear spectrum. The corresponding Gaussian distribution is $N(\mu^{Xlin}, \Sigma^{Xlin})$.

We assume that there is no correlation between the noise and speech signals. Since the noise signal and speech signal are additive in the linear domain, we get the equation:

$$X_{cep} = \Gamma^{-1} \log(e^{\Gamma S_{cep}} + e^{\Gamma N_{cep}}), \quad (1)$$

where Γ is the cosine transform, e is the function that calculates the exponential of each vector component, and log is the function that calculates the logarithm of each vector component. Now, we want to find the distribution of X_{cep}, when distributions of N_{cep} and S_{cep} are defined as Gaussian distributions. Since the cosine transform is a linear transform, the distribution after transformation is still a Gaussian distribution; distributions of $S_{lg}(=\Gamma S_{cep})$ and $N_{lg}(=\Gamma N_{cep})$ become Gaussian distributions. Thus the mean and covariance values of the distribution of $S_{lg}(=\Gamma S_{cep})$ after cosine transformation are easily calculated as follows.

$$\mu^{S_{lg}} = \Gamma \mu^{S_{cep}}. \quad (2)$$

$$\Sigma^{S_{lg}} = \Gamma \Sigma^{S_{cep}} \Gamma^{t}. \quad (3)$$

Thus the mean and covariance values of the distribution of $N_{lg}(=\Gamma N_{cep})$ are also calculated using the same formulations.

Since after the exponential transformation, the Gaussian distribution is a log normal distribution, the distributions of $S_{lin}(=e^{\Gamma S_{cep}})$ and $N_{lin}(=e^{\Gamma N_{cep}})$ are log normal distributions. Thus the mean and covariance values of the distribution of $S_{lin}(=e^{\Gamma S_{cep}})$ are calculated as follows.

$$\mu_u^{S_{lin}} = \exp(\mu_u^{S_{lg}} + \frac{\sigma_{uu}^{S_{lg}}}{2}). \quad (4)$$

$$\sigma_{uv}^{S_{lin}} = \mu_u^{S_{lin}} \mu_v^{S_{lin}} (\exp(\sigma_{uv}^{S_{lg}}) - 1). \quad (5)$$

Since $X_{lin}(=e^{\Gamma S_{cep}} + e^{\Gamma N_{cep}})$ is the sum of $e^{\Gamma S_{cep}}$ and $e^{\Gamma N_{cep}}$, the distribution of $X_{lin}(=e^{\Gamma S_{cep}} + e^{\Gamma N_{cep}})$ is obtained by convoluting the two distributions of $e^{\Gamma S_{cep}}$ and $e^{\Gamma N_{cep}}$. However, it is difficult to obtain a real distribution for X_{lin}. If we assume that X_{lin} can be approximated to a log normal distribution, the distribution of $X_{lg} = \log(e^{\Gamma S_{cep}} + e^{\Gamma N_{cep}})$ is a Gaussian distribution and its mean and covariance can easily be obtained from the mean and covariance of the distribution of $e^{\Gamma S_{cep}} + e^{\Gamma N_{cep}}$.

From this knowledge, the following equations are obtained.

$$\mu_u^{X_{lg}} = \log(\mu_u^{S_{lin}} + \mu_u^{N_{lin}}) - \frac{1}{2}\log(\frac{\sigma_{uu}^{S_{lin}} + \sigma_{uu}^{N_{lin}}}{(\mu_u^{S_{lin}} + \mu_u^{N_{lin}})^2} + 1). \quad (6)$$

$$\sigma_{uv}^{X_{lg}} = \log(\frac{\sigma_{uv}^{S_{lin}} + \sigma_{uv}^{N_{lin}}}{(\mu_u^{S_{lin}} + \mu_u^{N_{lin}})(\mu_v^{S_{lin}} + \mu_v^{N_{lin}})} + 1). \quad (7)$$

$$\mu^{X_{cep}} = \Gamma^{-1} \mu^{X_{lg}}. \quad (8)$$

$$\Sigma^{X_{cep}} = \Gamma^{-1} \Sigma^{X_{lg}} \Gamma^{-1t}. \quad (9)$$

Here, t: transpose; u,v: parameter indices, where $0 \le u,v \le p$ and $p+1$ is number of coefficients in the spectrum domain.

3. EXTENDED HMM COMPOSITION

We extended HMM composition to deal with additive noise and multiplicative distortion. We modeled speech signal in general noisy conditions as shown in Figure 2. Here, almost all variables are defined in the linear spectral domain, so the suffix of spectrum is omitted in the following notations. Speech signal S is produced by speech HMMs. Noise signal N is produced by a noise HMM. S and

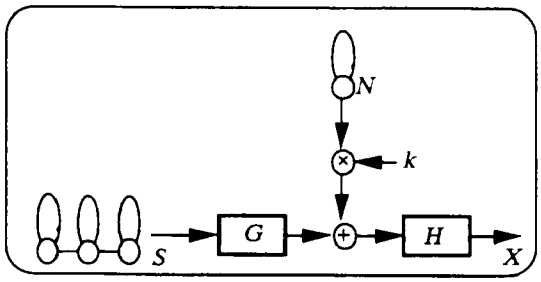

Figure 2: Model for producing noisy and distorted speech.

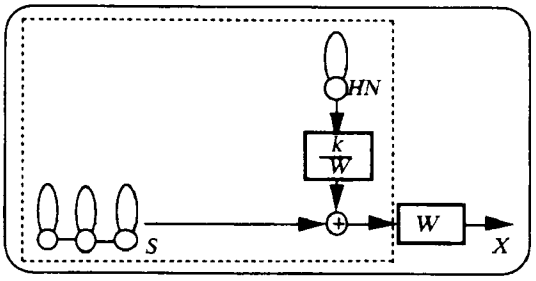

Figure 3: Converted model for producing noisy and distorted speech.

N are defined in the linear power spectral domain. S is first multiplied by multiplicative distortion G corresponding to distortion before noise is added. Then additive noise N is added to speech signal GS. Since the S/N ratio in evaluation is different from that in training, to control this factor, N is multiplied by coefficient k. Finally, the speech signal is multiplied by multiplicative distortion H corresponding to distortion after noise is added, for example microphone line distortion, microphone distortion, etc. Using these notations, the final noisy and distorted speech signal is $X = H(GS + kN) = HGS + kHN$. By setting $W = HG$, we obtain $X = WS + kHN = W(S + kHN/W)$, so the basic noisy speech model can be converted into the model shown in Figure 3. We suppose that the variables W and k are deterministic, not random. 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, and so we extended ordinary HMM composition so that composed HMMs could become a function of k and W. A set of phoneme HMMs producing X is modeled by composing the kHN/W HMM, the S HMMs and W. 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. To maximize $P(O|M(k,W))$, extended HMM composition basically uses an iterative procedure. In each iteration, the maximization process consists of both k and W estimation.

The value of k is estimated using the parallel model method, in which 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 values of W is estimated using the EM algorithm.

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

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 method estimates cepstral bias by using the EM algorithm [6].

4. INCORPORATING POWER VARIANCE

We found that when we used cep0 variance in the HMM composition process, the recognition rates degraded significantly, so until now, we have not used the variance of cep0 in HMM composition or extended HMM composition processes. That is the variance of cep0 was set to zero. The reason of this degradation might be as follows. Equations (6) and (7) approximate a complex distribution in the linear spectrum domain as a log normal distribution. When the log normal distribution is approximated using cep0 variance, its approximation accuracy degrades, since the variance of cep0 is much bigger than those of the other cepstra.

To avoid this problem, we approximate a complex distribution as the sum of several log normal distributions. To do this, the Gaussian distribution of cep0 is quantized as shown in Figure 4. Since we use diagonal Gaussian distributions in the cepstral domain as output probability in HMMs, using these quantized power values, a Gaussian distribution of cepstrum is divided into several Gaussian distributions with different cep0 means. (In the original HMM composition, variance of cep0 was set to zero; i.e., the distribution was quantized only at the original cep0 mean value.) The probability densities of cep0 at quantized points are normalized so that their sum is 1.0. These values are used as weights of the divided distributions. After quantization, extended HMM composition is performed for each divided Gaussian distribution.

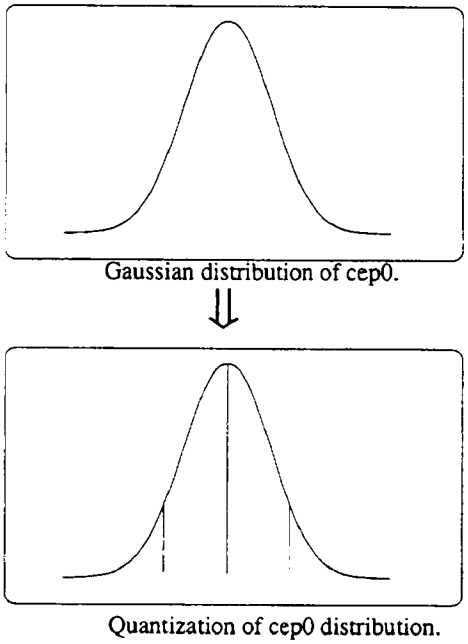

Figure 4: Approximation of cep0 distribution.

5. EXPERIMENTS

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. A16-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 universal speech HMM had one state with a 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, digit 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.

1785 digit strings uttered by 51 speakers were simultaneously recorded using 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 microphone in an anechoic room. In the experiment, speech data were made by adding this noise to clean speech data so that the S/N ratios became 12 and 18 dB.

Extended HMM compositions with and without cep0 variance were compared.

Table 1 shows recognition results for extended HMM composition with and without cep0 variance. n_q shows how many points were quantized on the cep0 distribution. We quantized the distribution of cep0 at three points ($\mu_0^{x_{cep}} - \sqrt{\sigma_{00}^{x_{cep}}}$, $\mu_0^{x_{cep}}$, $\mu_0^{x_{cep}} + \sqrt{\sigma_{00}^{x_{cep}}}$) and five points ($\mu_0^{x_{cep}} - 2\sqrt{\sigma_{00}^{x_{cep}}}$, $\mu_0^{x_{cep}} - \sqrt{\sigma_{00}^{x_{cep}}}$, $\mu_0^{x_{cep}}$, $\mu_0^{x_{cep}} + \sqrt{\sigma_{00}^{x_{cep}}}$, $\mu_0^{x_{cep}} + 2\sqrt{\sigma_{00}^{x_{cep}}}$) . The method using quantization at three points improved the string recognition rate for both computer room noise and car noise, compared with extended HMM composition.

We expected the method using quantization at five points to be more accurate than that at three points, but its results were the almost same or worse. This means that we do not need consider the power distribution far from the mean value.

6. CONCLUSION

This paper described the improvement of extended HMM composition by incorporating variance of the 0th order cepstrum. In this method, the distribution of the 0th order cepstrum is quantized and then HMM composition is performed. This method was evaluated by a connected digit recognition experiment, in which four-digit strings were recognized when the number of digits was known. Two types of noise, computer room noise and car noise, were used and noisy and distorted speech data were made by adding these types of noise to speech data recorded using a boundary microphone. These results confirmed that incorporating cep0 variance (power variance) into HMM composition increases recognition accuracy.

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 the digit string data.

REFERENCES

[1] 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.

[2] 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.

[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] 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.

[5] Y. Minami and S. Furui, "Adaptation method based on HMM composition and EM algorithm", *Proc. Int. Conf. Acoust. Speech, Signal Process.*, May 1996, pp. 327-330.

[6] 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.

noise		noise in cmputer room		noise in car	
S/N ratio		12dB	18dB	12dB	18dB
without cep0 variance		60.3%	84.6%	90.1%	96.0%
with cep0 variance	n_q=3	61.4%	86.2%	91.1%	96.1%
	n_q=5	61.8%	85.9%	90.6%	95.7%

Table 1: Recognition results of extended HMM composition with and without incorporating cep0 variance