INTEGRATION OF HMM COMPOSITION AND A MICROPHONE ARRAY FOR OVERLAPPING SPEECH RECOGNITION

T. Takiguchi, M. Nishimura

IBM Research, Tokyo Research Laboratory,
1623-14, Shimotsuruma, Yamato-shi, Kanagawa, 242-8502, JAPAN

ABSTRACT

For hands-free speech recognition, it is desirable to acquire a speech signal of the highest quality possible, and to reduce the mismatch between the test utterance and the acoustic model. In this paper, we present a stochastic approach to integrate acoustic model adaptation and signal enhancement using a microphone array. With this method, it is possible to find speaker directions even at low SNRs. The enhanced speech is recognized by using composite HMMs which are able to represent the statistics of the overlapping speech. When the SNR of the target speaker's speech relative to the interfering speech was 0 dB, the composite-speech HMMs improved the recognition rate to 80.4%. Integrating composite HMMs and a microphone array further improved it to 94.2% - a very respectable improvement over the original 23.0% recognition rate for clean HMMs using a single microphone.

1. INTRODUCTION

Although large-vocabulary speech recognition systems perform remarkably well, recognition accuracy is degraded by the presence of interfering voices. A robust speech recognition method using composite HMMs has been proposed for countering additive noise in [1, 2], and some reports have already shown that the composite noisy HMMs represent the statistics of noisy speech well. We have applied the HMM composition method to overlapping speech recognition in [3]. The experimental results have shown that the composite HMM combining the target speech HMM and an interfering speech HMM can improve overlapping speech recognition.

In this paper, we attempt to strengthen the HMM composition method by using a microphone array. There have been several other microphone-array-based approaches for dealing with overlapping speech recognition (e.g. [4]). Although those microphone-array-based approaches enhance speech intelligibility, they do not deal with the mismatch between the beamformed data and the acoustic model. Conventional approaches for estimating the speaker direction focus on the short-term or long-term power of the speech signal. For overlapping speech recognition, the SNR of the speech may be approximately 0 dB. In such cases, it is difficult to find the speaker direction with those approaches. In this paper, we estimate the speaker direction from test data with GMMs (Gaussian Mixture Models) based on a maximum-likelihood criterion. With this method, it is possible to find speaker directions even at low SNRs. Then the next stage of processing uses HMM composition to reduce the mismatch between the beamformed data and the acoustic model.

First, we describe a method for estimating the speaker direction with an acoustic model. Following this, we describe a robust speech recognition method based on HMM composition for overlapping speech.

2. ESTIMATION OF THE SPEAKER DIRECTION

In the stochastic approach, the estimated word sequence $\hat{W}$ is given by

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, p(Y_{\hat{\theta}}|W)p(W), \tag{1}$$

where $Y_{\hat{\theta}}$ is the test data which is processed with a delay-and-sum beamformer (e.g. [5]), and $\hat{\theta}$ is the estimated speaker direction. The estimation of the speaker direction is handled in a maximum-likelihood framework

$$\begin{aligned}\hat{\theta} &= \underset{\theta}{\operatorname{argmax}}\, p(Y_\theta|M) \\ &= \underset{\theta}{\operatorname{argmax}} \sum_t \max\{\log p(y_\theta(t)|\lambda_{S1}^{(GMM)}), \\ &\qquad \log p(y_\theta(t)|\lambda_{S2}^{(GMM)}), \cdots\}\end{aligned} \tag{2}$$

where we find a GMM having the maximum likelihood for every frame. Now let $S1$, $S2$, $\cdots$, be sound sources in the target environment. The set of GMMs, M, is

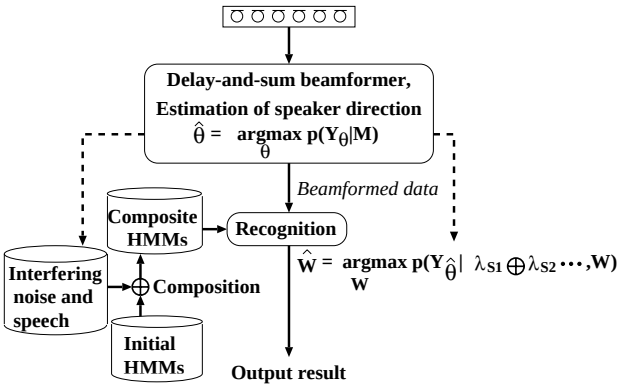

Figure 1: A robust speech recognition system using HMM composition and a microphone array

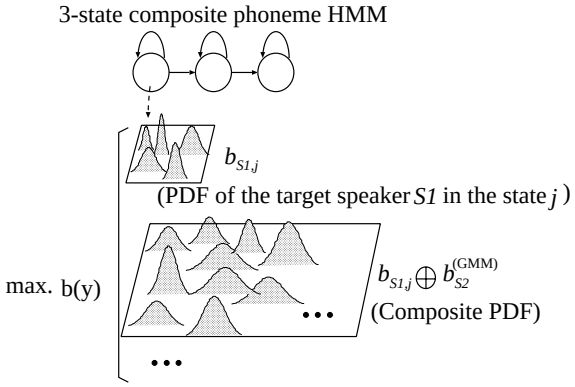

Figure 2: Composite HMM of the target speaker HMM and the interfering source GMMs.

given by

$$M = \{\lambda_{S1}^{(GMM)}, \lambda_{S2}^{(GMM)}, \cdots\}. \tag{3}$$

In equation (2), when the maximum likelihood for a frame is not calculated from the target GMM, the frame is rejected. When the percentage of rejected frames of the total frames is more than 70%, the estimated direction is rejected, and we use the last estimated direction.

Figure 1 shows a block diagram of the robust speech recognizer. Feature vectors are obtained by steering a beam to each direction, and the likelihood score for each direction is calculated with GMMs. Then, the direction having maximum likelihood is selected, and the beamformed signal is recognized by using composite HMMs.

3. HMM COMPOSITION FOR OVERLAPPING SPEECH

If the target signal $s1(t)$ and the interfering signals $s2(t)$, $\cdots$ are independent, the observed signal $o(t)$ is represented by

$$o(t) = s1(t) + s2(t) + \cdots. \tag{4}$$

To apply the HMM composition method to overlapping speech, the HMM parameters have to be transformed from the cepstral domain to the linear-spectral domain.

$$\begin{aligned} \lambda_{O_{cep}} \quad = \quad & \mathrm{Cos}^{-1}[\mathrm{Log}\{\ \mathrm{Exp}(\mathrm{Cos}(\lambda_{S1_{cep}})) \oplus \\ & k \cdot \mathrm{Exp}(\mathrm{Cos}(\lambda_{S2_{cep}})) \oplus \cdots\ \}]. \end{aligned} \tag{5}$$

Here λ denotes the set of HMM parameters, while the suffix *cep* represents the cepstral domain. The composition of HMMs is defined by the operator $\oplus$. The terms Cos, Log, and Exp are the cosine transform, logarithm transform, and exponential transform of the Gaussian probability density function, respectively. The SNR adjustment factor, k, is introduced to compensate for the mismatch of the signal level.

The overlapping phoneme HMM was made by using every possible combination of a target phoneme HMM and all possible interfering phoneme HMMs in [3], since the interfering speech is generally unknown. In new approach used this paper, to simplify the structure the overlapping phoneme HMM is given by the composition of the target phoneme HMM and the GMMs of the interfering sources. Figure 2 shows the structure of the composite HMM. The state j of the composite phoneme HMM has a set of observation probability density functions (PDFs):

$$B_j = \{b_{S1,j}, b_{S1,j} \oplus b_{S2}^{(GMM)}, \cdots\} \tag{6}$$

Here $b_{S1,j}$ is the PDF of the target speaker $S1$, and $b_{S1,j} \oplus b_{S2}^{(GMM)}$ is the composite PDF of $b_{S1,j}$ and the GMM of the interfering source $S2$. In each state, the PDF having the maximum likelihood is selected for every frame, and the likelihood is added to the total score.

The start and end points of the target speech are normally different from those of the interfering speech. We deal with the difference by adding a backward transition from the end state to the first state in the pause (silence) model.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Conditions

Word-recognition experiments were carried out on overlapping speech uttered by two males. Speaker dependent (SD) HMMs for the target speaker are trained by

Table 1: Word recognition rate [%] with a single microphone.

SNR	Clean HMM	Comp. HMM (0 dB)	Comp. HMM (5 dB)	Comp. HMM (10 dB)	Comp. HMM (15 dB)	Comp. HMM (20 dB)	Max. likelihood
0 dB	23.0	87.0	87.4	85.2	84.4	80.4	87.8
5 dB	37.4	89.0	91.0	91.8	92.4	92.0	92.0
15 dB	63.4	93.6	93.6	93.6	94.4	94.4	93.6

using 2620 words. The SD HMMs consist of 54 context-independent phonemes. Each HMM has three states and three self-loops, and each state has four Gaussian mixture components with diagonal covariance matrices. The interfering GMM is also trained by using 2620 words. The number of Gaussian mixture components is 256. For testing, we choose 500 words which are all different from the words used in training. The tests were carried out on 500-word recognition tasks. In the case of a single speaker, the recognition rate with the SD HMMs is 97.4%.

Microphone-array data is simulated considering only the time delay. Six microphones are uniformly spaced at 5 cm intervals, and the speech signal is sampled at 12 kHz. The target speaker and the interfering speaker are located at 45° and 135°, respectively.

4.2. Evaluation of HMM composition

The result for a single microphone is shown in table 1, where six sets of HMMs are used for evaluation. One set, the "Clean HMM," is the target SD model. The other sets are the composite HMMs. The "Comp. HMM (5 dB)" means the SNR is adjusted to 5 dB in equation (5).

At an SNR of 0 dB, the recognition rate with the clean HMMs is 23.0%. Using the composite HMMs (5 dB) increased the performance to 87.4%. This result is slightly better than that with "Comp. HMM (0 dB)". This is because the adjustment coefficient k in each test was different from that in HMM composition. The adjustment coefficient in the HMM composition is calculated by using all samples of training data.

Next, we employ the maximum-likelihood criterion to select the composite HMMs. After the calculation of the likelihood scores for each set of composite HMMs, the set of composite HMMs having the maximum likelihood score is selected. The recognition rate is shown in "Max. likelihood" of table 1. At the SNRs of 0 dB, 5 dB, and 15 dB, the recognition rates are 87.8%, 92.0%, and 93.6%, respectively. Comparing this result with the best one from each set of composite HMMs, the performance difference is relatively small. Therefore the maximum-likelihood criterion is effective in selecting composite HMMs.

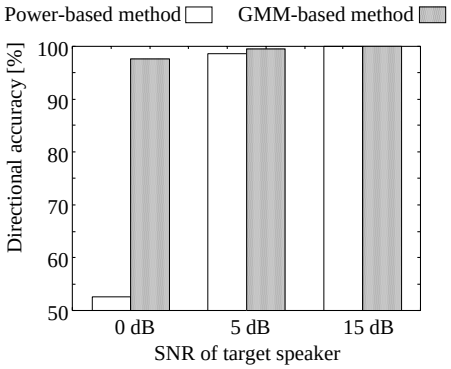

Figure 3: Directional accuracy for the target speaker. This figure compares the GMM-based method with the long-term-power-based method.

Table 2: Percentages of rejected words. In this table, [∗] shows the result for the interfering speaker.

SNR	Percentage of rejected words	Directional accuracy
0 dB	8.4% [20.0]	97.6% [90.5]
5 dB [-5]	3.0% [29.6]	99.6% [83.8]
15 dB [-15]	0.8% [50.2]	100% [79.5]

4.3. Evaluation of HMM composition and a microphone array

In this section, the test data is processed with a delay-and-sum beamformer. First, according to equation (2), the direction of the target speaker is estimated. Figure 3 shows the directional accuracy within a tolerance of 10°. With the GMM-based method, the directional accuracy is 97.6% at an SNR of 0 dB, 99.6% at an SNR of 5 dB, and 100% at an SNR of 15 dB. In comparison with the long-term-power-based method which finds the direction that maximizes the output power of the beamformer, there is not much difference in the performance, except for the SNR of 0 dB. With the power-based method, it is impossible to find the direction of the target speaker at low SNRs.

In equation (2), when the maximum likelihood for a frame is not calculated from the target GMM, the

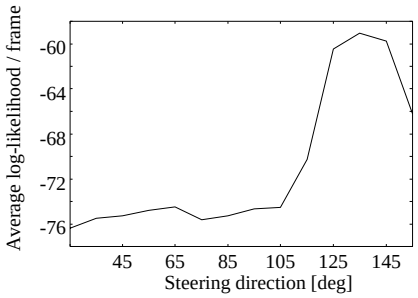

Figure 4: Average log-likelihood of interfering speech

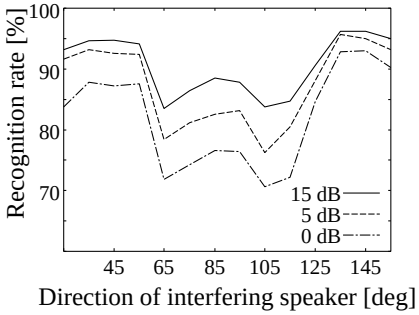

Figure 5: Word recognition rate for overlapping speech which is obtained by steering the beam to the estimated direction of the target speaker.

frame is rejected. When the percentage of rejected frames of the total frames is more than 70%, the estimated direction is rejected. Table 2 shows the percentages of the rejected words and the directional accuracy. It is seen from this table that the number of rejected words is increased when the SNR is low. At an SNR of 15 dB for the target speaker, the directional accuracy is 100.0%, and 0.8% of the test data is rejected. On the other hand, at an SNR of -15 dB for the interfering speaker, the directional accuracy is 79.5%, and 50.2% of the test data is rejected. When the SNR is very low, it is difficult to find the speaker direction using only the GMM-based method.

After the direction of the target speaker is estimated, the beamformed signal is recognized by using composite HMMs. Although the beamformed signal improves the SNR for the target speaker, the signal of the interfering speaker is distorted due to the frequency dependency of the directive patterns. In figure 4, we plot the average log-likelihood of the interfering speech versus the steering direction. The likelihood is calculated with the GMM of the interfering speaker. The interfering speaker is located at 135°. This figure shows that the likelihood is decreasing as the steering

Table 3: Word recognition rate [%] with a microphone array. [∗] shows the result with a single microphone.

SNR	Clean HMM	Comp. HMM (20 dB)
0 dB	51.4% [23.0]	94.2% [80.4]
5 dB	61.8% [37.4]	96.0% [92.0]
15 dB	73.6% [63.4]	96.4% [94.4]

direction is farther from the direction of the interfering speaker. To compensate for this, the beamformed signal for all possible directions are used for the GMM training (adapted GMM). Then the adapted GMMs and the HMMs of the target speaker are combined. Figure 5 shows the recognition rate for overlapping speech which is obtained by steering the beam to the estimated direction of the target speaker. By selecting the composite HMMs (for the direction of the interfering speaker) having the maximum-likelihood in figure 5, the recognition rate is improved from 51.4% (with the clean HMMs) to 94.2% at the SNR of 0 dB. In comparison with the performance of HMM composition shown in table 1, integrating composite HMMs and the microphone array improved it from 80.4% to 94.2% at the SNR of 0 dB, from 92.0% to 96.0% at the SNR of 5 dB, and from 94.4% to 96.4% at the SNR of 15 dB. These results are summarized in table 3.

5. CONCLUSION

This paper has presented a stochastic approach to integrate the acoustic model adaptation and signal enhancement with a microphone array. The experimental results show that it is possible to find the speaker direction even at low SNRs, and then the recognition rate for overlapping speech can be improved by using HMM composition and a microphone array.

6. REFERENCES

[1] M.J.F. Gales and S.J. Young, "An improved approach to the hidden Markov model decomposition of speech and noise," in *Proc. ICASSP,* pp. 233-236, 1992.

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

[3] T. Takiguchi and M. Nishimura, "Recognizing overlapping speech by using HMM composition," in *Proc. WESTPRAC,* pp. 89-92, 2000.

[4] P. Heracleous, T. Yamada, S. Nakamura, and K. Shikano, "Simultaneous recognition of multiple sound sources based on 3-D N-best search using a microphone array," in *Proc. Eurospeech,* pp. 69-72, 1999.

[5] T. Yamada, "Hands-free speech recognition using a microphone array," Doctor thesis, Nara Institute of Science and Technology, Japan, 1999.