The Seventh Western Pacific Regional Acoustics Conference
WESTPRAC VII
Kumamoto, Japan, 3-5 October 2000

RECOGNIZING OVERLAPPING SPEECH BY USING HMM COMPOSITION

Tetsuya TAKIGUCHI and Masafumi NISHIMURA

IBM Japan, Ltd., Tokyo Research Laboratory
1623-14 Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japan
Fax: +81 46 274 4282
E-mail: takigu@jp.ibm.com

ABSTRACT

This paper describes a robust speech recognition method based on HMM (Hidden Markov Model) composition for overlapping speech. The HMM composition method has already been proposed for additive noise, and some reports have shown that the composite noisy HMMs represent the statistics of noisy speech well. In this paper, we apply the HMM composition method to overlapping speech recognition. Since the observed signal is represented as the sum of the target speech and the interfering speech in the time domain, HMM composition should be effective for this purpose. Word-recognition experiments were carried out on overlapping speech uttered by two males. At a signal-to-noise-ratio (SNR) of 0 dB, the recognition rate with clean-speech HMMs was 28.0%. Using the composite speech HMMs increased the performance to 90.0%. At an SNR of 15 dB, the recognition rate improved from 67.5% to 95.0%.
KEYWORDS: HMM composition, multiple speakers, overlapping speech

INTRODUCTION

Although large-vocabulary speech recognition systems perform remarkably well, recognition accuracy is degraded by the presence of interfering speech. There have been various approaches to dealing with the problem of simultaneous speech from multiple speakers. These approaches are classified as speech-segregation-based systems (e.g. [1, 2, 3]) and microphone-array-based systems (e.g. [4, 5]). Most of them are still unable to resolve the problems of spectral distortion, reverberation and so on.

A robust speech recognition method using HMM composition to counter additive noise has been proposed [6, 7], and some reports have already shown that the composite noisy HMMs represent the statistics of noisy speech well (e.g. [8, 9]). In [10, 11], the use of HMM composition was proposed for countering both additive noise and reverberation. However, none of the above reports have discussed interfering speech. Robust speech recognition in the presence of interfering speech remains a serious problem.

In this paper, we apply the HMM composition method to overlapping speech recognition. Since the observed signal is represented as the sum of the target speech and the interfering speech in the time domain, HMM composition should also be effective for this purpose. The next section describes a robust speech recognition method based on HMM composition for overlapping speech. Following this, the performance for overlapping speech uttered by two males is shown.

HMM COMPOSITION FOR OVERLAPPING SPEECH

On the assumption that target signal $s1(t)$ and interfering signal $s2(t)$ are independent, the observed signal $o(t)$ is represented by

$$o(t) = s1(t) + s2(t). \tag{1}$$

This relation is preserved in the linear-spectral domain as follows:

$$O(\omega; m) = S1(\omega; m) + S2(\omega; m), \tag{2}$$

where $O(\omega; m), S1(\omega; m)$, and $S2(\omega; m)$ are short-term linear spectra in the analysis window m. The HMM composition executes the addition in the model domain instead of in the time domain. The parameters for speech recognition are represented by the cepstrum. To apply the HMM composition to overlapping speech, the HMM parameters have to be transformed from the cepstral domain to the linear-spectral domain. Therefore, the overlapping speech HMMs are given by

$$\lambda_{O_{cep}} = \mathrm{Cos}^{-1}[\mathrm{Log}\{\ \mathrm{Exp}(\mathrm{Cos}(\lambda_{S1_{cep}})) \oplus\ k \cdot \mathrm{Exp}(\mathrm{Cos}(\lambda_{S2_{cep}}))\ \}]. \tag{3}$$

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 $\oplus$ in this paper. The terms Cos, Log, and Exp are the cosine transform, logarithm transform, and exponential transform of the Gaussian probability density function, respectively.

The levels of the target and interfering signals are generally different in training and testing. Therefore, we will have to compensate for the mismatch of the levels. The conventional approach is to introduce an adjustment factor k in the linear-spectral domain. As shown in equation (3), the interfering speech HMMs are multiplied by the adjustment factor in this paper.

The overlapping phoneme HMM is made using every possible combination of a target phoneme HMM and the interfering phoneme HMMs, since the interfering speech is generally unknown. Figure 1 shows the overlapping phoneme HMM composed of the target phoneme HMM /o/ and

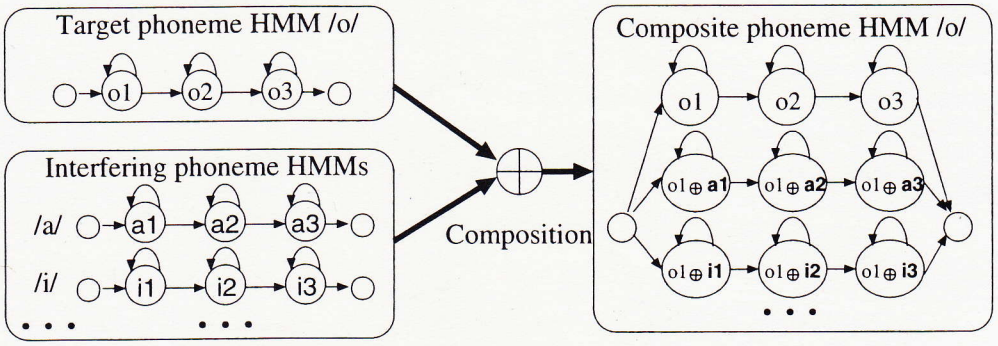

Fig.1 Example of overlapping speech HMM, where each phoneme HMM for target and interfering speech has three states. The above simplified structure of the overlapping speech HMM is used in this paper.

interfering phoneme HMMs, where each phoneme HMM for target and interfering speech has three states. The structure of the overlapping speech HMM is given by the Cartesian product of the component HMMs. In this paper, to simplify the structure, the following equation is defined:

(number of states for an overlapping phoneme)
= (number of states for a target phoneme) × (total number of interfering phonemes).

The start and end points of the target speech are 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.

The HMM recognizer decodes overlapping speech on a trellis diagram according to maximizing the log-likelihood. The decoded path will find an optimal combination of target and interfering speech.

EXPERIMENTS AND RESULTS

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 using 2620 words. The SD HMMs consist of 54 context-independent phonemes. The interfering HMMs are also trained by using 2620 words. Each HMM has three states and three self-loops, and each state has four Gaussian mixture components with diagonal covariance matrices. For testing, we choose 200 words which are different from the words used in training. The tests were carried out on 200-word recognition tasks. In the case of a single speaker, the recognition rate with the SD HMMs is 97.0%.

The speech signal is sampled at 12 kHz and windowed with a 32-msec Hamming window every 8 msec. Then FFT is used to calculate the 16th-order MFCCs (mel-frequency cepstral coefficients) and the power. In this paper, we use only the 16th-order MFCCs without their first order differentials. The power term is only used to adjust the SNR in HMM composition.

Experimental Results Recognition experiments were conducted to evaluate the performance of HMM composition at various SNRs. Table 1 shows the recognition rates for the target speaker at SNRs of 0 dB, 5 dB, and 15 dB, where four sets of HMMs are used. One set, "Clean HMM," is the clean HMMs for the target speaker. The other sets are the composite HMMs of the target HMMs and the interfering HMMs. At an SNR of 0 dB, the recognition rate with the clean HMMs is 28.0%. Using the composite HMMs increased the performance to 90.0%. At an SNR of 15 dB, the recognition rate improved from 67.5% to 95.0%. At an SNR of 5 dB, the recognition rate with "composite HMM (15 dB)" is slightly better than that with "composite HMM (5 dB)". This is because the adjustment coefficient k for each word is different from that used in the HMM composition. The adjustment coefficient in the HMM composition is calculated by using all the samples of training data. On the other hand, the adjustment coefficients for testing are calculated for each word. When the same adjustment coefficient is used in HMM composition and for each word, the recognition rate is 90.5% at an SNR of 0 dB, 94.0% at an SNR of 5 dB, and 95.5% at

Table.1 Word-recognition rates at various SNRs

SNR	Clean HMM	Composite HMM (0 dB)	Composite HMM (5 dB)	Composite HMM (15 dB)
0 dB	28.0%	90.0%	89.5%	89.5%
5 dB	49.0%	90.5%	92.5%	93.5%
15 dB	67.5%	91.0%	92.5%	95.0%

an SNR of 15 dB. This performance is slightly better than that shown in table 1. However, it is difficult to exactly adjust the power term, especially when the SNR is low. The other approaches may be necessary in real environments.

Next, we select the composite HMMs having the maximum likelihood. The recognition rate is improved to 93.0% at an SNR of 0 dB, 94.0% at an SNR of 5 dB, and 93.0% at an SNR of 15 dB. This performance is better than that shown in table 1, except for an SNR of 15 dB. There are several possible sources for the performance degradation. One might be the phoneme-connection to the interfering speech, where every combination of phonemes is included. Further improvements would be necessary for practical use.

CONCLUSION

This paper investigated the performance of HMM composition for recognition of overlapping speech uttered by two males. The experimental results show that the composite HMM of the target speech HMM and the interfering speech HMM can improve the performance of speech recognition in the presence of interfering speech. In future work, we will investigate the performance of HMM composition and separation [12, 13] for overlapping speech in noisy and reverberant environments.

REFERENCES

1. G. J. Brown and M. Cooke, "Computational auditory scene analysis," in *Computer Speech and Language*, 8, pp. 297-336, 1994.
2. P. D. Green, M. P. Cooke, and M. D. Crawford, "Auditory scene analysis and hidden Markov model recognition of speech in noise," in *Proc. ICASSP*, pp. 401-404, 1995.
3. H. G. Okuno, T. Nakatani, and T. Kawabata, "Understanding three simultaneous speeches," in *Proc. IJCAI*, pp. 30-35, 1997.
4. A. J. Bell and T. J. Sejnowski, "Blind separation and blind deconvolution," in *Proc. ICASSP*, pp. 3415-3418, 1995.
5. 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.
6. 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.
7. F. Martin, K. Shikano, and Y. Minami, "Recognition of noisy speech by composition of hidden Markov models," in *Proc. EUROSPEECH*, pp. 1031-1034, 1993.
8. Y. Minami and S. Furui, "Adaptation method based on HMM composition and EM algorithm," in *Proc. ICASSP*, pp. 327-330, 1996.
9. J. Hung, J. Shen, and L. Lee, "Improved parallel model combination techniques with split Gaussian mixtures for speech recognition under noisy conditions," in *Proc. ICASSP*, pp. 437-440, 1999.
10. S. Nakamura, T. Takiguchi, and K. Shikano, "Noise and room acoustics distorted speech recognition by HMM composition," in *Proc. ICASSP*, pp. 69-72, 1996.
11. T. Takiguchi, S. Nakamura, Q. Huo, and K. Shikano, "Adaptation of model parameters by HMM decomposition in noisy reverberant environments," in *Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 155-158, 1997.
12. T. Takiguchi, S. Nakamura, and K. Shikano, "Speech recognition for a distant moving speaker based on HMM composition and separation," in *ICASSP*, 2000.
13. T. Takiguchi, S. Nakamura, and K. Shikano, "HMM-Separation-Based speech recognition for a distant moving speaker," *IEEE Transactions on Speech and Audio Processing*, in printing.