論文 / 著書情報
Article / Book Information

Title	Speaker Recognition Using HMM Composition in Noisy Environments
Authors	Tomoko Matsui, Tomohiko Kanno, Sadaoki Furui
Citation	Eurospeech 95, Vol. 1, No. , pp. 621-624
Pub. date	1995, 9

Eurospeech '95

SPEAKER RECOGNITION USING HMM COMPOSITION IN NOISY ENVIRONMENTS

Tomoko Matsui†, Tomohito Kanno‡ and Sadaoki Furui†‡

†NTT Human Interface Laboratories, 3-9-11, Midori-cho, Musashino-shi, Tokyo, Japan
‡Tokyo Institute of Technology, 2-12-1, Ookayama, Meguro-ku, Tokyo, Japan

ABSTRACT

This paper investigates a speaker recognition method that is robust against background noise. In noisy environments, one important issue is how to create a model for each speaker so as to compensate for noise. The method described here is based on hidden Markov model (HMM) composition by the noise-and-voice (NOVO) transform. The HMM composition combines a speaker HMM and a noise-source HMM into a noise-added speaker HMM with a particular signal-to-noise ratio (SNR). Since it is difficult to measure the SNR exactly for non-stationary noise, this method creates several noise-added speaker HMMs with various SNRs. The HMM that has the highest likelihood value for the input speech is selected, and a speaker decision is made using this likelihood value.

Experimental application of this method to text-independent speaker identification and verification in various kinds of noisy environments demonstrated considerable improvement in speaker recognition for speech utterances of male speakers.

1 INTRODUCTION

While many speaker recognition systems demonstrate high performance, most of them are based on speech recorded under low-noise conditions. Hidden Markov models (HMMs) are often used for characterizing the speaker's voice in these systems; however, HMMs are sensitive to incoming noise, often resulting in sharply decreased speaker recognition.

Using HMM composition to create noise-added speech HMMs for noisy speech is a powerful technique for a wide range of tasks in speech and speaker recognition in noisy environments. Martin's noise-and-voice (NOVO) transform [1] is an effective HMM composition technique for speech recognition. Gales reported another effective noise compensation technique: parallel model combination [2]. Rose's HMM composition technique for speaker identification integrates signal-background models using the expectation-maximization algorithm [3]. However, these techniques require knowing the SNR, which is difficult to measure exactly for non-stationary noise.

The method we propose in this paper uses HMM composition by NOVO transform to cope with the problem of unknown SNRs.

2 METHOD

During the training phase (Figure 1), a speaker HMM is created for each speaker by using training speech recorded under noise-free conditions, while the noise-source HMM is created using noise signals recorded under the same conditions as those for recognition. Several noise-added speaker HMMs are then made for each speaker by combining the speaker HMM and the noise-source HMM, using the NOVO transform with various SNRs.

In the NOVO transform [1], since speaker and noise-source HMMs are defined in the cepstrum domain, and speech and noise signals are additive in the linear spectrum domain, the normal distributions defined for the mixture components of the states of the HMMs in the cepstrum domain are transformed into log-normal distributions in the linear spectrum domain and summed according to the SNR. The distributions obtained in the linear spectrum domain are finally converted back into the cepstrum domain.

During the recognition phase, for each noise-added speaker HMM of each speaker, the likelihood value of the input speech matching the HMM is calculated and the maximum value is used as the likelihood value of the speaker (shown in Figure 2).

3 EXPERIMENTAL CONDITIONS

The speech database consisted of sentence data uttered by ten male speakers. The sentences were selected from phonetically balanced sentences [4] and read. The speech was recorded during five sessions over ten months in the same noise-free room and using the same microphone for all speakers for all sessions. The average duration of each sentence was 4.2 s. The sampling rate was 12 kHz. The cepstral coefficients were calculated by LPC analysis with an order of 16, a frame period of 8 ms, and a frame length of 32 ms. Ten sentences from one session were used to create each speaker HMM. The noise database consisted

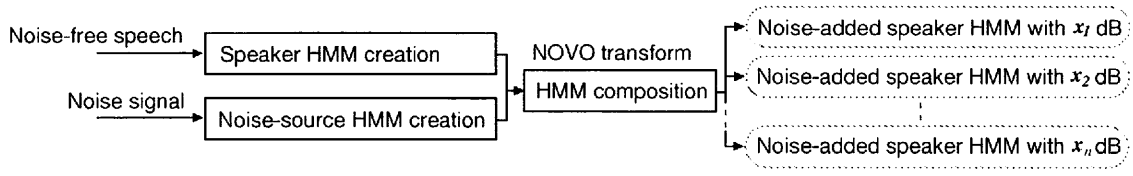

Figure 1. Training procedure (x_i : SNR).

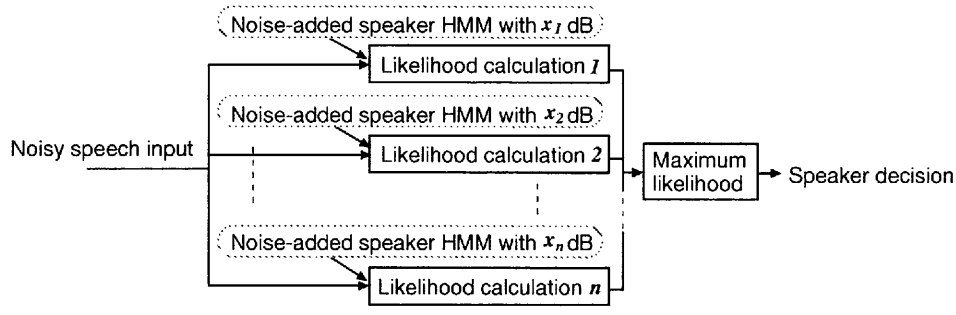

Figure 2. Recognition procedure (x_i : SNR).

Table 1. Speaker identification error rates (%) using test utterances with various SNRs.

Input SNR	0 dB			6 dB			12 dB			18 dB		
Model	noise-free	SNR-known	multi-SNRs	noise-free	SNR-known	multi-SNRs	noise-free	SNR-known	multi-SNRs	noise-free	SNR-known	multi-SNRs
Source:												
Elevator	73.5	43.5	49.0	49.5	10.5	8.5	25.0	5.0	5.5	8.5	3.5	3.5
Crowd	45.5	15.5	10.0	12.0	5.5	5.0	6.0	3.0	4.0	4.0	3.0	3.0
Computer	83.5	25.0	10.0	57.0	11.5	6.0	24.5	5.5	4.0	5.5	3.0	1.5
Telephone	31.0	13.5	11.5	13.0	7.5	6.5	5.5	5.5	5.5	4.5	4.0	4.0
Car	13.5	4.0	4.5	6.5	3.5	4.0	4.5	4.0	4.0	4.0	4.0	4.0
Exhibition	53.5	12.5	9.5	18.5	6.0	4.5	6.5	3.5	3.5	5.0	2.0	3.0
Average	50.1	19.0	15.8	26.1	7.4	5.8	12.0	4.4	4.4	5.2	3.2	3.2

Table 2. Speaker verification error rates (%) using test utterances with various SNRs.

Input SNR	0 dB			6 dB			12 dB			18 dB		
Model	noise-free	SNR-known	multi-SNRs	noise-free	SNR-known	multi-SNRs	noise-free	SNR-known	multi-SNRs	noise-free	SNR-known	multi-SNRs
Source:												
Elevator	36.6	18.8	19.5	16.1	4.1	3.4	4.0	1.1	1.1	0.8	0.3	0.3
Crowd	12.9	3.2	2.9	2.1	0.7	0.9	0.8	0.3	0.2	0.2	0.2	0.1
Computer	24.0	4.5	4.6	9.2	2.3	2.3	2.6	0.9	1.0	0.8	0.4	0.3
Telephone	9.9	1.8	2.2	2.1	0.6	0.6	0.5	0.2	0.1	0.1	0.1	0.1
Car	0.9	0.4	0.4	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Exhibition	9.9	3.2	2.9	2.3	1.3	1.2	0.5	0.5	0.4	0.2	0.2	0.2
Average	15.7	5.3	5.4	5.3	1.5	1.4	1.4	0.5	0.5	0.3	0.2	0.2

of six different kinds of noise sources recorded in (1) an elevator lobby, (2) a crowd, (3) a computer room, (4) a telephone booth, (5) a running car, and (6) an exhibition hall. To create each noise-source HMM, noise signals about two-minutes long were extracted and used. For testing, five sentences from the other four sessions that differed from the training sentences were used individually. The test data was made by adding noise signals (different parts from those used for training) to each speech to provide average SNRs of 0, 6, 12, and 18 dB. One-state, 64-mixture continuous HMMs were used as the speaker HMMs, and

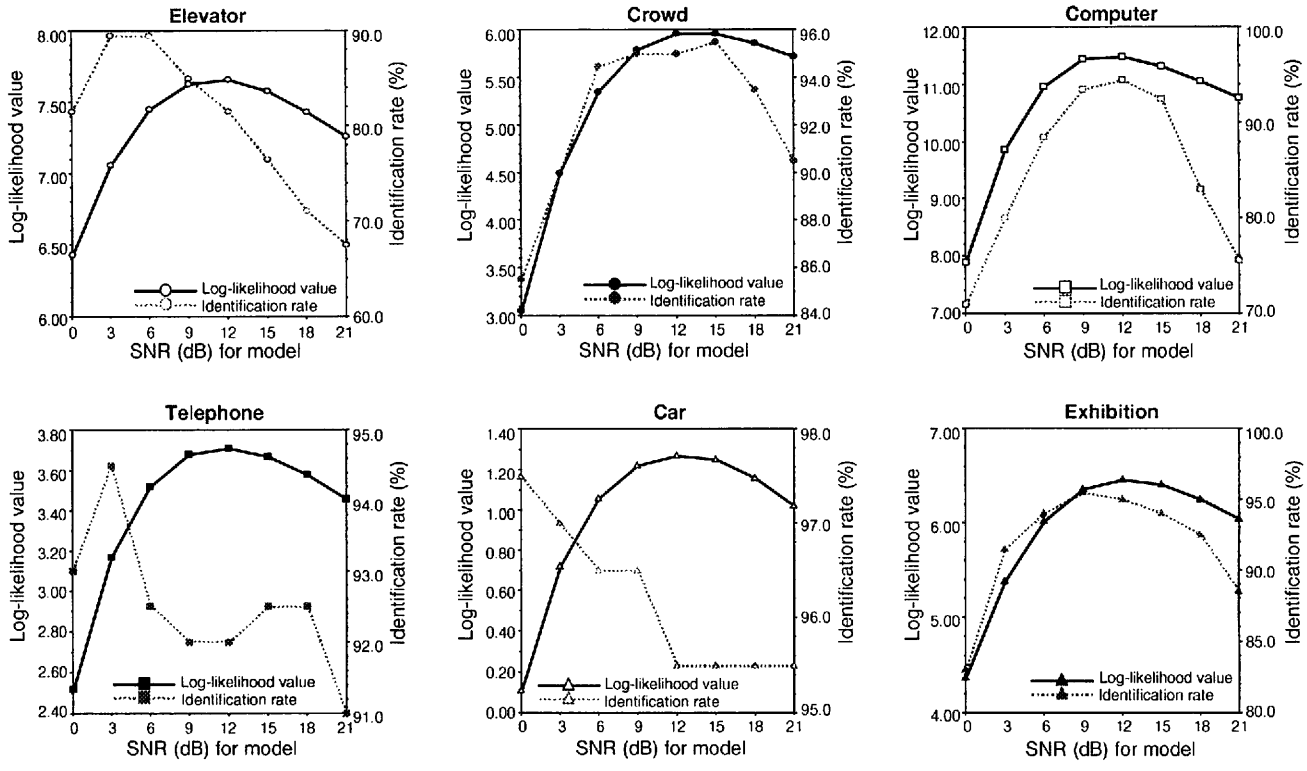

Figure 3. Log-likelihood values and identification rates (%) using test utterances with a 6-dB input SNR for noise-added speaker HMMs with various SNRs.

one-state, one-mixture continuous HMMs were used as the noise-source HMMs.

In the experiments, the following three methods were compared:

1. Speaker HMM for noise-free speech used directly [noise-free].
2. Noise-added speaker HMM used with a known SNR [SNR-known].
3. Multiple noise-added speaker HMMs with SNRs ranging from 0 to 18 dB in 3-dB steps used with unknown SNRs [multi-SNRs].

The identification and verification error rates for text-independent speaker recognition [5] were used for evaluation. The verification was performed using one speaker as the registered customer and the other nine speakers as impostors, rotating through all speakers and then averaging the results. The likelihood normalization method based on a posteriori probability [6] was used in the verification experiments. The threshold was set a posteriori to equalize the probability of false acceptance and false rejection.

4 RESULTS

Table 1 lists the identification error rates and Table 2 lists the verification error rates. When using speaker HMMs for noise-free speech and test data without noise signals, the identification error rate was 5.0% and the verification error rate was 0.1%.

These results show that the HMM composition method for "SNR-known" and "multi-SNRs" reduces the error rates for all noise sources. The "multi-SNRs" method performed as well as or better than the "SNR-known" method for all SNRs. High recognition performance was obtained using "multi-SNRs" even when the SNR was low and unknown.

5 DISCUSSION

5.1 Relation between the likelihood value and recognition rate

In our method, the maximum likelihood value is selected from those of noise-added speaker HMMs with various SNRs. Let us consider the relation between likelihood values and identification rates.

Figure 3 shows log-likelihood values and speaker identification rates as a function of the SNR of noise-added speaker HMMs (model SNR). The SNR of the input speech was fixed at 6 dB, and the values averaged over all input utterances at all sessions for each model SNR condition were plotted in this figure. For all types of noise, the models with an SNR of 12 dB – 6 dB higher than the actual input SNR – showed the highest log-likelihood values. The actual model SNR that has the maximum likelihood values varies

according to the input speech. Therefore, the identification rates when using multi-SNR models for a 6-dB input SNR, as shown in Table 2, are much higher than the values plotted in Figure 3.

The curves of the log-likelihood values and speaker identification rates correspond well for "Crowd", "Computer", and "Exhibition" noises. On the other hand, the model SNRs with the best identification rates are much lower than those with the maximum log-likelihood values for the other three noises. A supplementary experiment showed that the former group of noises is spectrally flatter and more stationary than the latter group. Further analysis and study is needed to establish a method to select the best model SNR conditions, especially for nonstationary noises. It would also be interesting to determine why the model SNRs corresponding to the highest identification rates are 6 dB higher on average than the actual input SNR.

5.2 Noise-source HMM structure

Although one-state, one-mixture HMMs were used as noise-source HMMs in the above experiments, they may not adequately represent noise that includes multiple noise-sources such as background voices and footsteps. Therefore, we performed speaker identification experiments using one-state, five-mixture HMMs as noise-source HMMs for the "Elevator" noise and using test utterances recorded in the same session as that for training.

Table 3 compares the identification error rates when one-state, five-mixture HMMs were used with those when using one-mixture HMMs. The SNR for input speech was known, and noise-added speaker HMMs with the same SNR as that for input speech were used. Only when the SNR was 0 dB, did one-mixture HMMs perform better than five-mixture HMMs; with all other SNRs, the five-mixture HMMs performed better. These results suggest that the recognition performance presented in this paper could be further improved by using more complex noise-source HMMs according to the complexity of noise characteristics.

Table 3. Speaker identification error rates (%) for noise-added speaker HMMs composed of various structures of noise-source HMMs for the "Elevator" noise using test utterances that were recorded in the same session as that for training.

Input SNR	0 dB	6 dB	12 dB	18 dB
Noise-source HMM:				
one-state, five-mixture	48	12	2	0
one-state, one-mixture	42	16	4	2

6 CONCLUSIONS

HMM composition by NOVO transform is effective for speaker recognition in noisy environments. Our method obtained high recognition performance even when the SNR was unknown.

Although noise-added speaker HMMs with SNRs ranging from 0 to 18 dB in 3-dB steps were used and the maximum likelihood value was used for speaker decision, further study is needed to improve the method of adjusting the step size according to the likelihood values of noise-added speaker HMMs and of selecting the likelihood value for speaker decision so as to obtain the best performance.

7 ACKNOWLEDGMENT

The authors wish to acknowledge the members of the Furui Research Laboratory of NTT Human Interface Laboratories for their valuable and stimulating discussions.

REFERENCES

[1] F. Martin, K. Shikano, and Y. Minami. "Recognition of Noisy Speech by Composition of Hidden Markov Models", Proc. Eurospeech, Berlin, pp. II-1031-1034, 1993.

[2] M.J.F. Gales and S.J. Young, "HMM recognition in noise using parallel model combination", Proc. Eurospeech, Berlin, pp. II-837-840, 1993.

[3] R.C. Rose, E.M. Hofstetter, and D.A. Reynolds. "Integrated Models of Signal and Background with Application to Speaker Identification in Noise." IEEE Trans. Speech and Audio Processing, Vol. 2, No. 2, pp.245-257, 1994.

[4] H. Kuwabara et al., "Construction of ATR Japanese Speech Database as a Research Tool," ATR Technical Report, TR-I-0086, 1989.

[5] T. Matsui and S. Furui, "Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMMs," Proc. ICASSP, San Francisco, pp. II-157-160, 1992.

[6] T. Matsui and S. Furui, "Similarity normalization method for speaker verification based on a posteriori probability," Proc. ESCA workshop on automatic speaker recognition identification and verification, Martigny, pp.59-62, 1994.