NOISE AND ROOM ACOUSTICS DISTORTED SPEECH RECOGNITION BY HMM COMPOSITION

Satoshi NAKAMURA, Tetsuya TAKIGUCHI, Kiyohiro SHIKANO

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama Ikoma, Nara, 630-01 JAPAN
e-mail: nakamura@is.aist-nara.ac.jp

ABSTRACT

This paper presents a robust speech recognition method based on the HMM composition for the noisy room acoustics distorted speech. The method realizes an improved user interface such as the user is not encumbered by microphone equipments. The proposed HMM composition is obtained by naturally extending the HMM composition method of an additive noise to that of the convolutional room acoustics distortion. The HMM composition is conducted by 2 steps: 1)Composition of HMMs of a speech and acoustical transfer function in cepstrum domain, 2)Composition of distorted speech and noise HMMs in linear spectral domain. The speaker dependent/independent word recognition experiments are carried out using the speech database contaminated by the additive noise and convolutional room acoustics distortion. The evaluation experiments are also conducted for unknown testing sound source positions. These results clarified the effectiveness of the proposed method.

1. INTRODUCTION

In spite of recent advances, the speech recognition technology did not reach to the practical use in the real world. The reason is that the advances are almost achieved in the error reduction of the clean speech recognition.

A key issue way to the practical use is a development of a recognition technology of noise and room acoustics distorted speech. This technology will especially take an important role on recognition of distant-talking speech.

The accuracy of speaker independent speech recognition is made a remarkable progress by the arrival of stochastic modeling of speech, HMM, and its training algorithms. Although the HMM brought a high recognition accuracy, a speaker must be equipped a close-talking microphone. If the speaker inputs his speech from the distance or through a telephone channel, the accuracy will be drastically degraded by the influences of the room acoustic or telephone channel distortion and environment noises. Therefore we have to overcome the two problems such as,

- Additive noise
- Convolutional distortion

Many works are presented to solve these problems. These approach are summarized as follows:

- Speech Enhancement.
- Model Modification.

As for the speech enhancement approach, the spectral subtraction method for an additive noise and the cepstral mean normalization method for a convolutional noise had been proposed and confirmed their effectiveness[1, 2]. As for the model modification approach, the conventional multi-template approach, and model adaptation approach[8] and the model (de-)composition approach[3, 4, 6] had been proposed. Among these approaches the HMM composition approach is the most promising, because the HMM for the noisy speech can be easily generated by composing the speech HMMs and the noise HMM which trained during noise period. The paper[4, 6] showed the composed noisy HMM outperforms very high accuracy.

In this paper, we apply the HMM composition to the recognition of the speech which is contaminated by not only an additive noise but also the room acoustics distortion. If the components are independent each other and additive, HMM composition can be adopted. The noise and speech are independent and additive in linear spectral domain. While the transfer function and speech are convolutional in linear spectral domain, they are independent and additive in cepstral domain. Thus the HMM composition is applicable for noise and room acoustics distorted speech. Some studies have been already presented for the problem of spectral tilt compensation of speech with a noise and channel distortion[5, 7]. This paper addresses the compensation not only the spectral tilt but also room acoustical transfer function.

This paper presents the HMM composition algorithm for the noise and room acoustics distorted speech and evaluates its effectiveness by the recognition experiments of the distant-talking speech in a noisy room, where the speech is suffered from the noise and room acoustics distortion.

2. HMM COMPOSITION

On the assumption that the speech and noise signal are independent, the observed signal is represented by

$$O(t) = S(t) + N(t)$$

The conventional approach estimates noise statistics during the noise period and recognizes an input noisy speech by using the noise added reference patterns. The HMM composition executes addition in HMM parameter domain instead of the addition in signal domain. Since the signal

0-7803-3192-3/96 $5.00©1996 IEEE

level is generally different between training and testing, an adjustment factor k is introduced. Thus the observed signal is represented by

$$O(t) = S(t) + kN(t)$$

where $O(t), S(t)$ and $N(t)$ are the observed noisy signal, speech signal and noise signal, respectively. Since this relation is preserved in linear spectral domain, we regard $O(t), S(t), N(t)$ as short time linear spectra whose analysis window starts at time t from now on.

Generally, parameters for speech recognition are represented by the cepstrum. The parameters have to be transformed to linear domain as an addition of the speech and noise holds[4, 6].

As for a convolutional distortion, the observed distorted spectrum is represented by

$$O(t) = H(t) \cdot S(t)$$

where $H(t)$ is a transfer function from the sound source to the microphone. $H(t)$ is a function of time t since the sound source may move. The multiplication can be converted to sum in cepstral domain as,

$$O_{cep}(t) = H_{cep}(t) + S_{cep}(t)$$

where, $O_{cep}(t), H_{cep}(t)$ and $S_{cep}(t)$ are cepstra for the observed signal, acoustic transfer function and speech signal, respectively. Therefore the observed signal is represented by

$$O(t) = exp(\mathcal{F}^{-1}(S_{cep}(t) + H_{cep}(t))) + kN(t) \qquad (1)$$

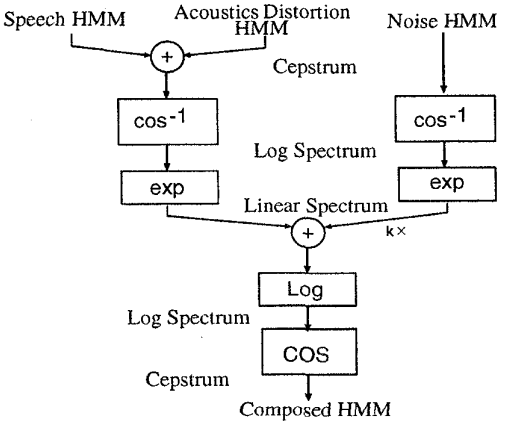

Figure 1. Block diagram of HMM Composition

This procedure is summarized in Fig.1. The cosine transform, inverse cosine transform, exponential transform and log transform are conducted on HMM parameters.

The procedure is as follows:

1. Estimate HMMs of the speech, noise and acoustical transfer function in cepstral domain.
2. Compose HMMs of the speech and acoustical transfer function in cepstral domain.

$$\mu_{(cep\_SH)} = \mu_{(cep\_S)} + \mu_{(cep\_H)}$$

$$\Sigma_{(cep\_SH)} = \Sigma_{(cep\_S)} + \Sigma_{(cep\_H)}$$

Here, $\mu_{(cep\_S)}, \Sigma_{(cep\_S)}, \mu_{(cep\_H)}, \Sigma_{(cep\_N)}, \mu_{(cep\_SH)}$ and $\Sigma_{(cep\_SH)}$ are a mean vector and covariance matrix of HMMs of the speech, acoustical transfer function, and composed HMMs in cepstral domain, respectively.

3. Cosine transform of each Gaussian pdf of HMMs.

$$\mu_{(log\_SH)} = \Gamma \cdot \mu_{(cep\_SH)}$$

$$\Sigma_{(log\_SH)} = \Gamma \cdot \Sigma_{(cep\_SH)} \cdot \Gamma^{-1}$$

Here, Γ is a cosine transform matrix, $\mu_{(log_SH)}$ and $\Sigma_{(log_SH)}$ are a mean vector and covariance matrix of Gaussian pdf in log power spectral domain, respectively.

4. Exponential transform to linear spectral domain. The normal random vectors obtained by exponential transform, $Z = exp^Y$, has log normal distribution. A mean and covariance are given by,

$$\mu_{(lin\_SH),i} = exp\{\mu_{(log\_SH),i} + \frac{\sigma^2_{(log\_SH),ii}}{2}\}$$

$$\sigma^2_{(lin\_SH),ij} = \mu_{(log\_SH),i} \cdot \mu_{(log\_SH),j} \cdot \{\exp(\sigma^2_{(log\_SH),ij} - 1)\}$$

Here, $\mu_{(lin\_SH)}$ and $\Sigma_{(lin\_SH)}$ are a mean vector and covariance matrix in linear power spectral domain.

5. Compose two distributions according to the equation(1).

$$\mu_{(lin\_SHN)} = \mu_{(lin\_SH)} + k \cdot \mu_{(lin\_N)}$$

$$\Sigma_{(lin\_SHN)} = \Sigma_{(lin\_SH)} + k^2 \cdot \Sigma_{(lin\_N)}$$

Here, $\mu_{(lin\_N)}, \Sigma_{(lin\_N)}, \mu_{(lin\_SHN)}$ and $\Sigma_{(lin\_SHN)}$ are a mean vector and covariance matrix of the noise and composed model in linear power spectral domain, respectively.

6. Log transform of composed HMM.

$$\mu_{(log\_SHN),i} = \log \mu_{(lin\_SHN),i} - \frac{1}{2}\{\frac{\sigma^2_{(lin\_SHN),ij}}{\mu_{(lin\_SHN),i} \cdot \mu_{(lin\_SHN),i}} + 1\}$$

$$\sigma^2_{(log\_SHN),ij} = \log\{\frac{\sigma^2_{(lin\_SHN),ij}}{\mu_{(lin\_SHN),i} \cdot \mu_{(lin\_SHN),j}} + 1\}$$

7. Inverse cosine transform to cepstral domain.

$$\mu_{(cep\_SHN)} = \Gamma^{-1} \cdot \mu_{(log\_SHN)}$$

$$\Sigma_{(cep\_SHN)} = \Gamma^{-1} \cdot \Sigma_{(log\_SHN)} \cdot \Gamma$$

The HMM recognizer decodes the observed signal on a trellis diagram according to maximize the log likelihood. Decoded path will bring a optimal combination of a speech, transfer function and noise.

3. EXPERIMENTS

The speech recognition experiments are carried out to investigate the effectiveness of the proposed method. The evaluation of the length of the impulse response and of unknown testing sound source positions is also conducted.

We conducted recognition experiments of the degraded speech uttered in a noisy room. Fig.2 shows the room used in the experiment. We measured 9 transfer functions from 9 positions to the microphone. The former five positions, $h_1, \cdots, h_5$ are used for the model composition and the latter four positions, $p_1, \cdots, p_4$ are used for the recognition tests. The transfer functions are measured by the sweep method. The length of the original impulse response was 180msec(L=180msec). The test data are simulated by the linear convolution of speech corpus and the measured transfer function. The Fig.3 shows the cepstral coefficients of acoustical transfer functions from several training positions. This differences will cause the degradation on speech recognition.

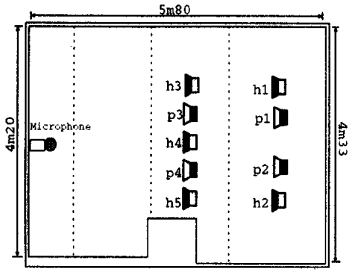

Figure 2. Room environment

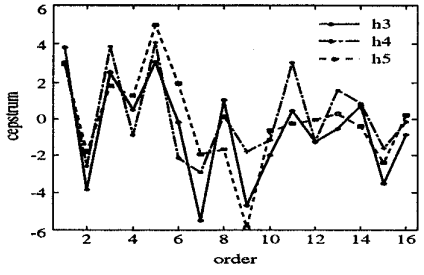

Figure 3. Differences of cepstral coefficients

Speech corpus for evaluation is ATR Japanese speech database Set-A and ASJ continuous speech database. This database contains word and sentence utterances by announcers. The recognition algorithm is based on 256 tied-mixture diagonal covariance HMMs. The HMM has 5 states and 3 self loops. The Context independent 54 phone models are trained by 2620 words. The other 500 words are used for testing. We prepared speaker dependent(SD) and independent(SI) HMMs. The SD HMMs are trained by 1 male and the SI HMMs are trained by 64 speakers. The experiments are carried out by using one male speaker used for SD training. The noise data is collected in a computer room and added to the clean speech data or acoustics distorted data as the SNR is 15dB.

Speech signal is sampled in 12kHz and windowed by 32msec Hamming window every 8msec. Then FFT is used to calculate 16-order MFCCs and a power. In the recognition, a power term is not used because it is only necessary to adjust the SNR in the HMM composition.

We assigned one state for the noise HMM and 5 states for the HMM of the acoustical transfer function. Fig.4 shows the structure of the HMM of the acoustical transfer function. Each state directly corresponds to one of the training positions, $h_1, \cdots, h_5$. The single Gaussian pdf is used for these HMMs.

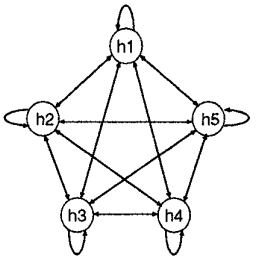

Figure 4. Ergodic HMM of acoustical transfer function

The speech recognition experiments are conducted. The points to be investigated are:

- Improvements of a recognition rate for the noisy distorted speech.
- Evaluation of a speaker dependent(**SD**) and speaker independent(**SI**) speech recognition performance.
- Performance for an unknown sound source position.

The spectral analysis for speech recognition is based on short time windowing. The multiplication of short time signal spectra and the transfer function is equivalent to the periodic convolution in time domain. However, an actual distorted signal is made by the linear convolution. Since the proposed HMM composition of the signal and room acoustical transfer function only realizes the periodic convolution, the composed HMM can't model a actual room acoustics distorted signal accurately. The difference of using periodic and linear convolution will be large according to the length of impulse response.

In this paper, the transfer function in cepstral domain is obtained by subtracting the cepstrum coefficients of original speech from those of convolutioned speech. Although this transfer function will be affected by the predecessor samples, HMM would be able to model the variation by covariance matrix, Σ, of Gaussian pdf.

The results are listed in Table.1. The results are summarized as follows:

- The HMM composition improves the speech recognition rate both for noisy speech and distorted speech, from 29.8% to 92.4% and from 86.4 to 92.0%, respectively. This means HMM composition successfully models the noisy or distorted speech.

Table 1. Model specification and speaker dependent/independent recognition rates[%]

Input	HMM-S	HMM-SN	HMM-SH(μ)	HMM-SH(μ, Σ)	HMM-SHN(μ)	HMM-SHN(μ, Σ)
Noise Composition	×	○	×	×	○	○
Acoustics Composition	×	×	○	○	○	○
Clean speech	96.6/92.6	-	-	-	-	-
Noisy speech	29.8/47.4	92.4/83.2	-	-	-	-
Distorted speech	86.4/74.9	-	92.0/77.3	87.8/70.4	-	-
Noisy Distorted Speech	15.0/24.5	72.3/55.4	-	-	76.5/59.1	66.6/42.4

- For the noisy distorted speech, the improvement is obtained from 15.0% to 76.5%. Although the improvement is about 61.5%, the performance is supposed to be still insufficient. SD recognition rate for one known training position using HMM trained by the noisy distorted speech is 82.2%. The recognition rate for the same data using HMM-SHN is 77.2%.
- The effect of the covariance matrix, Σ, of HMM of the acoustical transfer function is disappointing. This is because the variation is too large than expected and distributed dependently on predecessor speech characteristics.

Next, we evaluate the performance of the model for the unknown testing sound source positions. Table.2 shows average SD recognition rates for the known training and unknown testing positions. It is confirmed that the degradation between the training and testing sound source positions is relatively small for all composed HMM. The Fig.5 shows the euclidian distances and recognition rates by using each positions' acoustical transfer function for unknown testing position p_4. This table indicates a closest position results best performance and difference between the using acoustical transfer function of the closest position and using the ergodic HMM of the acoustical transfer function is quite small in this experiment.

Table 2. Recognition rates for known/unknown positions[%]

Input	HMM-S	HMM-SH(μ)	HMM-SH(μ, Σ)
Distorted speech	86.4/86.1	92.0/91.7	87.8/86.9

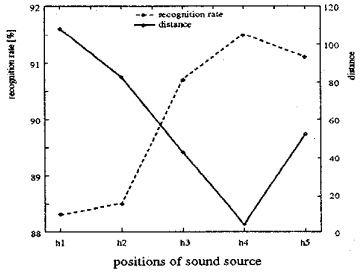

Figure 5. Recognition rates and distances

We have also examined the effect of a length of the impulse response. The several distorted signals are artificially made by L=180,100 and 32msec impulse responses by cutting out the original one. The Table.3 shows the results by HMM-SH. Although the simulated impulse response is not exist in real room, it is observed that the effects decrease according to the length of the impulse response.

Table 3. SD/SI recognition rates vs. the length of impulse response[%]

Input	180msec	100msec	32msec
Distorted speech	87.7/70.4	91.2/78.6	95.4/88.2

4. CONCLUSION

This paper presents a novel method which realizes an robust speech recognition of the noisy and room acoustics distorted speech. The method is based on the composition of HMMs of the speech, noise and acoustical transfer function. The experiments indicate that the proposed method improves the speaker dependent 500 Japanese word recognition rates from 15% to 76.5% for the noisy distorted speech. This improvement is less than expected by the improvements using either two composed HMM, HMM-SN or HMM-SH. The further improvement of the composed HMM for noisy distorted speech would be needed as a future work. The distorted speech from the unknown testing sound source positions is also evaluated. The degradation is found to be very small. It is also found that the variance of HMM of the acoustical transfer function is not able to compensate the difference between the linear convolution and periodic convolution.

REFERENCES

[1] S.F.Boll,"Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE,ASSP-27,No.2,1979

[2] A.Acero,*Acoustical and Environmental Robustness in Automatic Speech Recognition.* Ph.D Dissertation, ECE Department, CMU, Sep.1990

[3] A.P.Varga,R.K.Moore,"Hidden Markov Model Decomposition of Speech and Noise", ICASSP90, pp.845-848,1990

[4] M.J.F.Gales,S.J.Young,"An improved Approach to the Hidden Morkov Model Decomposition of Speech and Noise",ICASSP92, pp.233-236,1992

[5] M.J.F.Gales,S.J.Young,"PMC for Speech Recognition in Additive and Convolutional Noise",CUED-F-INFENG-TR154, 12, 1993

[6] F.Martin,K.Shikano,Y.Minami,"Recognition of Noisy Speech by Composition of Hidden Markov Models",EUROSPEECH93,pp.1031-1034,1993

[7] Y.Minami,S.Furui,"A Maximum Likelihood Procedure for a Universal Adaptation Method Based on HMM Composition", ICASSP95,pp.129-132, 1995

[8] A.Sankar,C-H.Lee,"Robust Speech Recognition Based on Stochastic Matching",ICASSP95,pp.121-124,1995