4pSC23

Acoustical Society of America and Acoustical Society of Japan

Third Joint Meeting

Honolulu, Hawaii, 2-6 December 1996

HANDS-FREE SPEECH RECOGNITION BY A MICROPHONE ARRAY AND HMM COMPOSITION

Satoshi Nakamura, Takeshi Yamada, Tetsuya Takiguchi, Kiyohiro Shikano

Graduate School of Information Science,
Nara Institute of Science and Technology
nakamura@is.aist-nara.ac.jp

1 INTRODUCTION

In real environments, acoustic ambient noise causes severe performance degradation of speech recognizers. One way to solve this problem is to use a head-mounted microphone. However it is seriously troublesome for speakers to be encumbered by microphone equipments. In order to make full use of speech interface, it is very important to use a hands-free speech input. Moreover as a hands-free speech input interface, the system has to recognize speech from distant mobile speakers. The speech from distant mobile speakers is suffered from noises and room reverberation. This paper presents two novel approaches to these problems. The first approach uses a microphone array. A microphone array is able to make full use of spatial phase information. The second approach is based on stochastic modeling of the observation signal by composing HMMs for each information sources. This paper also shows the effectiveness through speech recognition experiments in the noisy and reverberant room.

2 MICROPHONE ARRAY

Many techniques have been proposed to realize robust and hands-free speech recognition [1, 2], but most of these techniques strongly depend on noise characteristics. They work effectively only under restricted conditions. In recent years, a speech enhancement technique by a microphone array has been studied for speech recognition[3, 4, 5, 6, 7]. A microphone array is composed of multiple microphones which are spatially arranged and the outputs of each microphone have the phase differences according to the position of sound sources. To obtain the enhanced speech signal, this technique principally utilizes these information and forms the directivity which is sensitive to a speaker direction. Therefore this technique works effectively in variously noisy environments. In case of applying a microphone array to speech recognition, it is extremely important to localize a speaker direction accurately. Recently, some speech recognition systems using a microphone array have been proposed [6, 7]. However, it may be insufficient to localize a speaker direction in low SNR conditions. This paper proposes robust speech recognition with speaker localization based on extracting a pitch harmonics by a microphone array[8].

2.1 DELAY-AND-SUM BEAM-FORMER

A block diagram of the SLAM(Speaker Localization by an Arrayed Microphone) system is shown in Fig1. The SLAM system is composed of a speaker localizer, a speech enhancer and a speech recognizer.

In this paper, the delay-and-sum beamformer is used as a microphone array signal

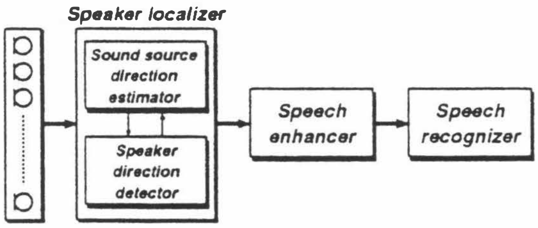

Figure 1: Block diagram of the SLAM system

processing. It is assumed that a plane wave comes from the direction θ to the equally spaced array composed of M microphones, where the plane wave is a complex sinusoidal signal in frequency f, and d denotes the distance between two adjacent microphones. The outputs of each microphone $x_1(t), \cdots, x_M(t)$ are given as follows:

$$x_i(t) = x_1\left(t - (i-1)\frac{d\cos\theta}{c}\right), \tag{1}$$

where c is the sound velocity and i is microphone index. Then the output of the delay-and-sum beam-former is given as follows:

$$\begin{aligned} y(t) &= \sum_{i=1}^{M} x_i\left(t + (i-1)\frac{d\cos\theta}{c}\right) \\ &= \sum_{i=1}^{M} x_i(t)\exp\left\{j2\pi f(i-1)\frac{d\cos\theta}{c}\right\} \end{aligned} \tag{2}$$

As a result of Eq. (2), a signal comes from the direction θ is M times as large, while signals come from different directions aren't enhanced. Therefore the directivity which is sensitive to the direction θ is formed.

2.2 SPEAKER LOCALIZATION ALGORITHM

An algorithm of the SLAM system is shown in Fig.2. The details of (A)(B)(C)(D) in Fig.2 are described as follows.

(A) Frequency analyzer In order to apply the delay-and-sum beam-former for broadband signals, the outputs of each microphone are divided into K frequency components. In Fig. 2, $x_1(n;m), \cdots, x_M(n;m)$ and $X_1(k;m), \cdots, X_M(k;m)$ denote the outputs of each microphone and the FFT of them. Where n, k, and m are sample index, frequency index, and frame index, respectively.

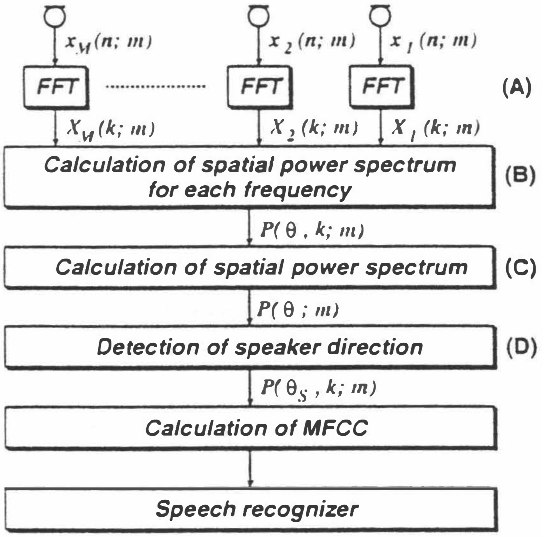

Figure 2: Algorithm of the SLAM system

(B) (C) Sound source direction estimator In order to estimate sound source directions $\theta_1, \cdots, \theta_\Gamma$, the spatial power spectrum which defined by Eq. (3) is calculated.

$$P(\theta;m) = \sum_{k=0}^{K-1} P(\theta,k;m), \theta = 0, 1, \cdots, 180, \tag{3}$$

where $P(\theta,k;m)$ is equivalent to the output power of the delay-and-sum beamformer, and is given as follows:

$$\begin{aligned} &P(\theta,k;m) \\ &= \left|\sum_{i=1}^{M} X_i(k;m)\exp\left\{j2\pi f_k(i-1)\frac{d\cos\theta}{c}\right\}\right|^2 \end{aligned} \tag{4}$$

where $\theta = 0, 1, \cdots, 180$ and f_k denotes a corresponding frequency to k. Γ sound source directions are obtained by detecting every peaks of directions on the spatial power spectrum.

(D) Speaker direction detector A speaker direction θ_S is detected from among the sound source directions estimated in (B)(C). As a result, a enhanced speech power spectrum is obtained as $P(\theta_S, k; m), k = 0, \cdots, K-1$.

A simple speaker localization algorithm is based on extracting the maximum power (SLAM-P). SLAM-P is represented as $\theta_S =$

$\mathrm{argmax}_{\theta_\gamma} P(\theta_\gamma; m)$, where θ_γ denotes one of Γ sound source directions. However this algorithm will be in trouble in low SNR conditions. In this paper, a speaker localization algorithm based on extracting a pitch harmonics (SLAM-H) is used to localize a speaker direction accurately in low SNR conditions. SLAM-H is represented as $\theta_S = \mathrm{argmax}_{\theta_\delta} P(\theta_\delta; m)$, where θ_δ denotes one of Δ sound source directions extracted a pitch harmonics. If $\Delta = 0$, then a speaker direction which detected previously is used for current frame.

2.3 EXPERIMENTS

The microphone array is a equally spaced array composed of 14 microphones, where the distance between two adjacent microphones is 2.83 cm. The speaker direction and the Gaussian noise source direction are at 90 degree and 40 degree. The outputs of each microphone are generated considering only the time differences. In real environments, the experimental room as shown in Fig.3 is used for recording. The reverberant time in this room is about 0.18 sec. Two loud speakers are substituted for a speaker and a Gaussian noise source.

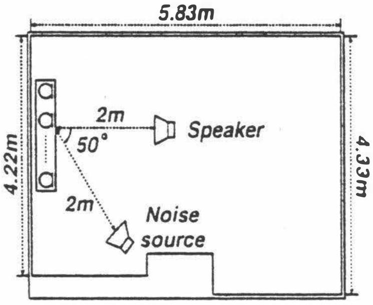

Figure 3: Experimental Room

The recognition algorithm is based on 256 Tied Mixture HMM. Speech signals are sampled at 12 kHz and windowed by the 32 ms Hamming window every 8 ms, and then calculated 16-order MFCCs and 16-order Δ MFCCs and a Δ Power. The recognition experiment is conducted for speaker dependent 500 words recognition. To evaluate the performance of the SLAM system, Word recognition Accuracy (WA) and Speaker Localization Accuracy (SLA) are used.(within ±3°)

The directional pattern obtained for 6 kHz band-limited Gaussian noise in computer simulation and real environments are shown in Fig.4. The gain for 40 degree in computer simulation and real environments are −21.7 dB and −17.3 dB.

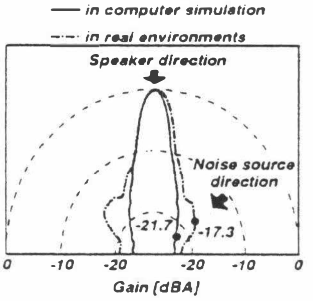

Figure 4: Directivitional Pattern

2.4 RECOGNITION RESULTS

The word recognition accuracy (WA) and speaker localization accuracy (SLA) are shown in Table 1. Delay-sum is the case that a speaker direction is known. This should be an upper bound of the performance of the SLAM system. On the other hand, SLAM-P and SLAM-H are the speaker direction unknown condition. Clean is the case that a Gaussian noise source isn't located (SNR 38 dB). This table confirms that the SLAM system with extraction of pitch harmonics attains the much higher speech recognition performance than that of a single microphone not only in computer simulation but also in real environments. There still remains degradation in the real environment. The de-reverberation of acoustical transfer function and normalization of microphone elements will be necessary.

Table 1: Word Recognition Accuracy and Source Localization Accuracy

Simulation	SNR [dB]				
	10		20		Clean
	WA	SLA	WA	SLA	WA
Single	27.4	—	74.8	—	97.2
Delay-sum	90.0	100.0	97.6	100.0	97.4
SLAM-P	29.8	24.1	83.0	56.6	—
SLAM-H	90.0	99.9	97.6	99.8	—
Real Env.	SNR [dB]				
	10		20		Clean
	WA	SLA	WA	SLA	WA
Single	11.4	—	53.4	—	85.6.
Delay-sum	64.2	100.0	82.0	100.0	85.8
SLAM-P	18.6	21.1	64.6	47.7	—
SLAM-H	55.6	80.3	81.2	98.7	—

3 HMM COMPOSITION

In this section HMM composition method for additive noise and room reverberation is introduced. Many works are presented to solve noise and reverberation problems from the points of speech enhancement and 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, 9]. As for the model modification approach, the conventional multi-template approach, and model adaptation approach[13] and the model (de-)composition approach[2, 10, 12] had been proposed. Among these approaches the HMM composition approach is 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 papers[10, 12] shows the composed noisy HMM outperforms very high accuracy. The papers[11, 14] also try to apply the method to telephone channel adaptation. 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 reverberation[15]

3.1 THEORY

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 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[10, 12].

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

$$O(t) = exp(\mathcal{F}(S_{cep}(t) + H_{cep}(t))) + kN(t) \quad (5)$$

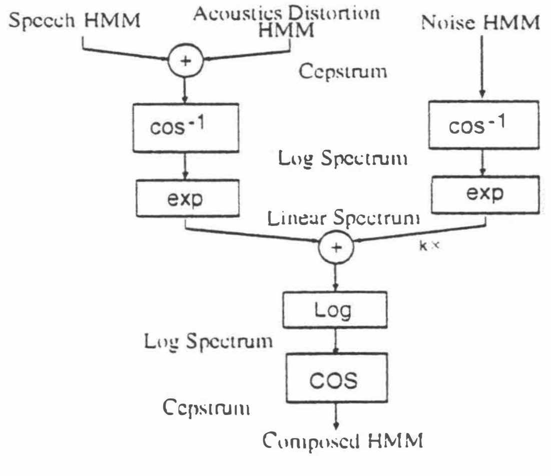

Figure 5: Block diagram of HMM Composition

This procedure is summarized in Fig.5. The cosine transform, inverse cosine transform, exponential transform and log transform are conducted on HMM parameters. 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.

3.2 EXPERIMENTS

The recognition experiments are conducted for degraded speech uttered in a same room used in section 2. Fig.6 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 test data are simulated by the linear convolution of speech corpus and the measured transfer function.

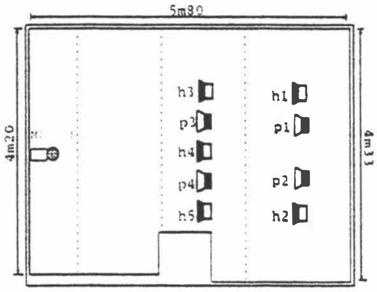

Figure 6: Room environment

Speech corpus and analysis conditions are the same as section 2. Speaker dependent(SD) and independent(SI) HMMs are prepared. The SI HMMs are trained by 64 speakers. The 500 words both in training data set and testing data set are used for recognition evaluation. The experiments are carried out by using two male and two female speakers. The noise data is collected in a computer room and added to the clean speech data or acoustically distorted data as the SNR is 15dB.

We assigned one state for the noise HMM and 5 states for the HMM of the acoustical transfer function. Fig.7 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.

The transfer function in cepstral domain is obtained by subtracting the cepstrum coefficients of original speech from those of convolutioned speech. However 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.2. The results clarify the effectiveness of the proposed method for noisy and reverberant speech recognition. The effect of a length of the impulse response is also examined. The several distorted signals are artificially made by L=180,100 and 32msec impulse responses by adjusting the original one. The Table.2 also shows those results by HMM composition.

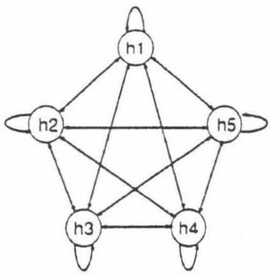

Figure 7: Ergodic HMM of acoustical transfer function

Next, we evaluate the performance of the model for the unknown testing sound source positions. Table.3 shows average SD recognition rates for the known training and unknown testing positions in 180msec reverberation. It is confirmed that the degradation between the training and testing sound source positions is relatively small for all composed HMM.

4 CONCLUSION

This paper describes two novel approaches to solve the problems of robust speech recognition in noisy reverberant room. Firstly, the approach based on SLAM, the speaker localization by using arrayed microphones, is presented. The experiments show that the delay-and-sum beam-former improves SNR by +20dB in 50° and the proposed localization algorithm estimates the speech source direction accurately. SLAM-H achieves the recognition rates close to those of speaker direction known case. Secondly, the HMM composition approach with a single microphone for an additive noise and room acoustical distortion is presented. The performance is evaluated by

Table 2: SD and SI word recognition rates for noisy distorted speech

Models	HMM-S		HMM-SN		HMM-SHN(μ)		HMM-SHN(μ, Σ)	
	SD	SI	SD	SI	SD	SI	SD	SI
Noise composition	×	×	○	○	○	○	○	○
Acoustics composition	×	×	×	×	○	○	○	○
180msec	4.8	18.7	59.5	53.5	67.2	57.2	55.2	45.4
100msec	9.9	22.0	76.2	65.7	83.6	66.7	79.7	59.2
32msec	14.4	21.9	76.5	65.4	87.0	68.4	86.5	65.5

Table 3: Results for known and unknown positions(L=180msec)

Models	HMM-S		HMM-SH(μ)		HMM-SH(μ, Σ)	
Acoustics composition	×	×	○	○	○	○
Distorted speech	known	unknown	known	unknown	known	unknown
	78.5	77.8	87.2	86.2	84.0	83.7

the noisy and reverberant speech. The results show that the proposed HMM composition improves the recognition accuracy from 4.8% to 67.2% for SD test by the HMM composition. These two approaches are quite effective to realize hands-free speech recognition in real adverse environments. Furthermore complementary combination of these approaches will bring more effective hands-free speech recognition system.

References

[1] S.F.Boll,"Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. ASSP-27, 4, pp. 113–120, April 1979.

[2] A.P.Varga,R.K.Moore, "Hidden Markov Model Decomposition of Speech and Noise", ICASSP90, S15b.10, pp. 845–848, April 1990.

[3] J.L.Flanagan, R.Mammone, G.W.Elko, "Autodirective microphone system for natural communication with speech recognizers", 4th DARPA Workshop on Speech and Natural Language, pp.4.8-4.13, 1991.2

[4] H.F.Silverman, S.E.Kirtman, J.E.Adcock, P.C.Meuse, "Experimental results for baseline speech recognition performance using input acquired from a linear microphone array", 5th DARPA Workshop on Speech and Natural Language, pp.285-290, 1992

[5] D.Van Compernolle, W.Ma, F.Xie, M.Van Diest, "Speech recognition in noisy environments with the aid of microphone arrays",Speech Communication, 9(5/6) pp.433-442, 1990

[6] Q.Lin, E.E.Jan, C.Che, B.de Vries, "System of microphone arrays and neural networks for robust speech recognition in multimedia environment", ICSLP94, S22-2, pp. 1247–1250, Sep. 1994.

[7] D.Giuliani, M.Matassoni, M.Omologo, P.Svaizer, "Hands-free continuous speech recognition in noisy environment using a four microphone array", ICASSP95, pp.860-863, 1995

[8] T.Yamada, S.Nakamura, K.Shikano, "Robust speech recognition with speaker localization by a microphone array",ICSLP96 1996,10

[9] A.Acero,*Acoustical and environmental robustness in automatic speech recognition*. Ph.D Dissertation, ECE Department, CMU, Sept.1990

[10] M.J.F.Gales, S.J.Young, "An improved approach to the hidden Markov model decomposition of Speech and Noise",ICASSP92, pp.233-236, 1992

[11] M.J.F.Gales, S.J.Young, "PMC for speech recognition in additive and convolutional noise", CUED-F-INFENG-TR154, 12, 1993

[12] F.Martin, K.Shikano, Y.Minami, "Recognition of noisy speech by composition of hidden Markov models", EUROSPEECH93, pp.1031-1034, 1993

[13] A.Sankar, C-H.Lee, "Robust speech recognition based on stochastic matching", ICASSP95, pp.121-124, 1995

[14] Y.Minami,S.Furui, "A maximum likelihood procedure for a universal adaptation method based on HMM composition", ICASSP95, pp.129-132, 1995

[15] S.Nakamura, T.Takiguchi, K.Shikano, "Noise and acoustics distorted speech recognition by HMM composition", ICASSP96, pp.69-72, 1996