ROBUST SPEECH RECOGNITION WITH SPEAKER LOCALIZATION BY A MICROPHONE ARRAY

Takeshi YAMADA, Satoshi NAKAMURA, and Kiyohiro SHIKANO

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, 630-01 Japan

ABSTRACT

This paper proposes robust speech recognition with Speaker Localization by a Arrayed Microphone (SLAM) to realize hands-free speech interface in noisy environments. In order to localize a speaker direction accurately in low SNR conditions, a speaker localization algorithm based on extracting a pitch harmonics is introduced. To evaluate the performance of the proposed system, speech recognition experiments are carried out both in computer simulation and real environments. These results show that the proposed system attains the much higher speech recognition performance than that of a single microphone not only in computer simulation but also in real environments.

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.

Many techniques have been proposed to realize robust and hands-free speech recognition [1][2][3], 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. 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 [4][5]. However, it is insufficient to localize a speaker direction in low SNR conditions. This paper proposes robust speech recognition with Speaker Localization by a Arrayed Microphone (SLAM) to realize hands-free speech interface in noisy environments. In order to localize a speaker direction accurately in low SNR conditions, a speaker localization algorithm based on extracting a pitch harmonics is introduced. To evaluate the performance of the proposed system, speech recognition experiments are carried out both in computer simulation and real environments [6][7].

2. SYSTEM OVERVIEW

2.1. Architecture of the SLAM System

A block diagram of the SLAM system is shown in Figure 1. The SLAM system is composed of a speaker localizer,

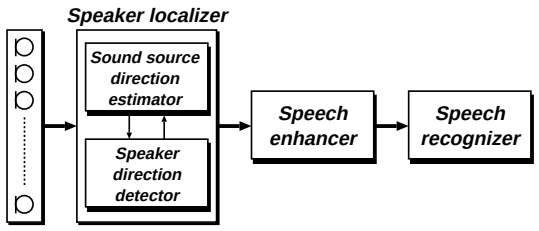

Figure 1: Block diagram of the SLAM system

a speech enhancer and a speech recognizer. The speaker localizer localizes a speaker direction under the condition that several noise sources exist. The speech enhancer obtains a enhanced speech signal by forming a directivity which is sensitive to a speaker direction.

2.2. Delay-and-sum Beamformer

In this paper, the delay-and-sum beamformer [8] is used as a microphone array signal processing. A principle of the delay-and-sum beamformer is shown in Figure 2. 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

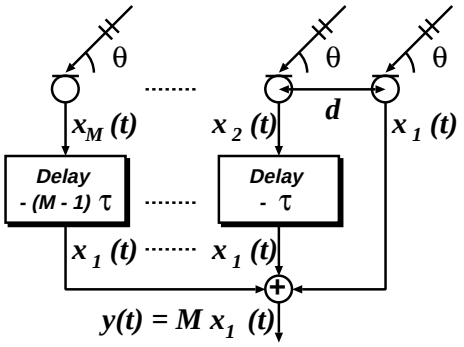

Figure 2: Delay-and-sum beamformer

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:

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

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

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

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.

2.3. Speaker Localization Algorithm

An algorithm of the SLAM system is shown in Figure 3. The details of (A)(B)(C)(D) in Figure 3 are described as follows.

(A) **Frequency analyzer** In order to apply the delay-and-sum beamformer for broadband signals, the outputs of each microphone are divided into K frequency components. In Figure 3, $x_1(n;m), \cdots, x_M(n;m)$ 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.

(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.

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

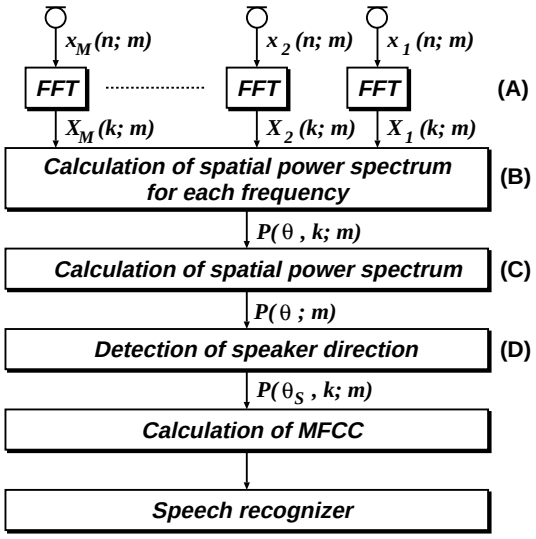

Figure 3: Algorithm of the SLAM system

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

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

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 [9]. 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.

3. EXPERIMENTS AND RESULTS

To evaluate the performance of the SLAM system, speech recognition experiments are carried out both in computer simulation and real environments.

3.1. Experimental Conditions

In computer simulation, sound sources and a microphone array are located as shown in Figure 4. The micro-

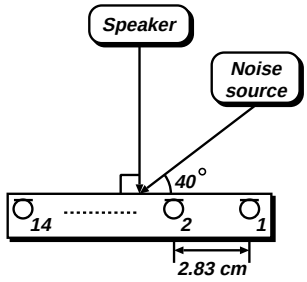

Figure 4: Sound sources and a microphone array

phone 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 Figure 5 is used for recording. The reverberant time in this room

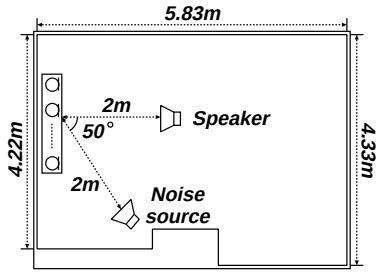

Figure 5: Experimental room

is about 0.18 sec. Two acoustic speakers are substituted for a speaker and a Gaussian noise source. The other experimental conditions are equally to those in computer simulation.

The recognition algorithm is based on 256 Tied Mixture HMM. As a speech corpus for speech recognition experiments, ATR Japanese speech database Set-A is used. MHT 2620 words are used for training context independent 54 phone models, another MHT 500 words for testing. 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.

To evaluate the performance of the SLAM system, Word recognition Accuracy (WA) and Speaker Localization Accuracy (SLA) are used. The SLA is defined as follows:

$$\mathrm{SLA} = \frac{\text{number of correct frames}}{\text{number of total frames}} \times 100[\%], \qquad (5)$$

where the number of correct frames means that of frames detected a correct speaker direction within 3 degree difference.

3.2. Directivities

The directivities obtained for 6 kHz band-limited Gaussian noise in computer simulation and real environments are shown in Figure 6. The gain for 40 degree in com-

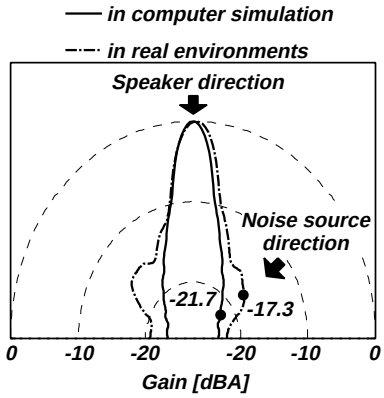

Figure 6: Directivities

puter simulation and real environments are -21.7 dB and -17.3 dB. In real environments, the acoustical channel distortion causes degradation of the gain.

3.3. Speech Recognition Results

The word recognition accuracy (WA) and speaker localization accuracy (SLA) in computer simulation are shown in Table 1. Delay-sum is the case that a speaker direction

	SNR [dB]				
	10		20		Clean
	WA	SLA	WA	SLA	WA
Single	27.4	—	74.8	—	97.2
Delay-sum	90.4	100.0	97.2	100.0	97.2
SLAM-P	28.8	24.6	81.8	57.4	—
SLAM-H	90.2	99.4	97.2	99.6	—

Table 1: Word recognition accuracy (WA) [%] and speaker localization accuracy (SLA) [%] (in computer simulation).

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). The results in computer simulation are summarized as follows:

1. Delay-sum achieved 97.2 % WA in SNR 20 dB and 90.4 % WA in SNR 10 dB. These results show that the Gaussian noise is almost removed in case of speaker direction known.
2. SLAM-H achieved 99.6 % SLA in SNR 20 dB and 99.4 % SLA even in SNR 10 dB, while SLAM-P is in trouble in low SNR conditions. As a result, SLAM-H attained 97.2 % WA in SNR 20 dB and 90.2 % WA in SNR 10 dB.

The WA and the SLA in real environments are shown in Table 2. The SNR in Clean is 32 dB. The results in real

	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	—
SLAM-P	18.6	20.7	62.6	47.8	—
SLAM-H	58.0	78.7	78.8	88.7	—

Table 2: Word recognition accuracy (WA) [%] and speaker localization accuracy (SLA) [%] (in real environments).

environments are summarized as follows:

1. The WA of Single in Clean is 85.6 % and the difference from the result in computer simulation is 11.6 %. This degradation is caused by the acoustical channel distortion.
2. The WA of Delay-sum in SNR 20 dB is 82.0 % and the difference from that in Clean is only 3.6 %. This result shows that the Gaussian noise is almost removed in case of speaker direction known.
3. SLAM-H achieved 88.7 % SLA in SNR 20 dB and 78.7 % SLA in SNR 10 dB, while SLAM-P is in trouble in low SNR conditions. As a result, SLAM-H attained 78.8 % WA in SNR 20 dB and 58.0 % WA in SNR 10 dB.

It is found that the SLAM system attains the much higher speech recognition performance than that of a single microphone not only in computer simulation but also in real environments.

4. CONCLUSION

This paper proposed robust speech recognition with Speaker Localization by a Arrayed Microphone (SLAM) to realize hands-free speech interface in noisy environments. In order to localize a speaker direction accurately in low SNR conditions, a speaker localization algorithm based on extracting a pitch harmonics was introduced.

To evaluate the performance of the proposed system, speech recognition experiments were carried out both in computer simulation and real environments. In computer simulation, the proposed system achieved 99.6 % Speaker Localization Accuracy (SLA) and 97.2 % Word recognition Accuracy (WA) in SNR 20 dB, and 99.4 % SLA and 90.2 % WA even in SNR 10 dB. In real environments, the proposed system achieved 88.7 % SLA and 78.8 % WA in SNR 20 dB, and 78.7 % SLA and 58.0 % WA in SNR 10 dB. These results show that the SLAM system attains the much higher speech recognition performance than that of a single microphone not only in computer simulation but also in real environments.

To improve the performance of the proposed system in real environments, it will be effective to compensate the acoustical channel distortion by an adaptive microphone array.

REFERENCES

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

[2] Lim, J.S. and Oppenheim, A.V., "All-pole modeling of degraded speech", IEEE Trans. ASSP-26, 6, pp. 197–210, June 1978.

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

[4] Qiguang Lin, Ea-Ee Jan, ChiWei Che, Bert de Vries, "System of Microphone Arrays and Neural Networks for Robust Speech Recognition in Multimedia Environment ", Proc. ICSLP94, S22-2, pp. 1247–1250, Sep. 1994.

[5] Giuliani, D., Omologo, M., and Svaizer, P., "Talker Localization and Speech Recognition Using a Microphone Array and a Cross-Powerspectrum Phase Analysis", Proc. ICSLP94, S22-1, pp. 1243–1246, Sep. 1994.

[6] Yamada, T., Nakamura, S., Shikano, K., "Speech Recognition with Speaker Localization by Microphone Array", Proc. ASJ meeting, 1-2-4, Sep. 1995 (in Japanese)

[7] Yamada, T., Nakamura, S., Shikano, K., "Robust Speech Recognition with Speaker Localization by a Microphone Array — Performance Evaluation in Real Environments —", Proc. ASJ meeting, 1-5-19, March 1996 (in Japanese)

[8] Pillai, S.U., "Array Signal Processing", Springer-Verlag, New York, 1989.

[9] Nakamura, S., Yamada, T., and Shikano, K., "Speech Recognition with Source Localization by Microphone Array", Proc. ASJ meeting, 1-5-8, March 1995 (in Japanese)