Real-time Sound Source Localization and Separation System and Its Application to Automatic Speech Recognition

Futoshi Asano, Masataka Goto, Katunobu Itou and Hideki Asoh

AIST (former ETL), Japan

asano@etl.go.jp

Abstract

A real-time sound localization/separation system for near-field sound sources was constructed and evaluated in a real office environment. As for the sound localization, the experimental results showed that the direction of the two sources was estimated with high accuracy while the range of the sources was estimated with moderate accuracy. As for the sound separation, a recognition rate of 70% for an on-line recognizer on a network and of 90% for an off-line recognizer were achieved, respectively.

1. Introduction

When using an automatic speech recognizer (ASR) in a real environment, its performance suffers deterioration due to environmental noise or signals from other sound sources. The authors previously proposed a method of sound source localization/separation using a microphone array [1]. In this method, assuming a dialog between a human and a mobile robot under conditions of sound interference, the sound localization/separation in the near-field range (the robot, the speaker and the interference are within a range of 1-2 m) was focused on. In the near-field, the wavefront becomes spherical and the range of the sources as well as the direction must be taken into account for accurate source localization/separation. As for the sound localization, by extending a high-resolution spatial spectrum estimator, MUSIC [2], to that for a 2-dimensional space, the range and the direction of sources can be estimated. Based on the localization results, a separation filter is designed using the minimum-variance beamformer (MVBF) [3].

In this paper, a real-time system which realizes this method was constructed using DSPs and was evaluated in a real office environment.

2. Method

2.1. Model of Signal

The Fourier transform of the microphone input, $X_m(k,t)$, is modeled as

$$\mathbf{x}(k,t) = [X_1(k,t), \cdots, X_M(k,t)]^T = \mathbf{A}_k \mathbf{s}(k,t) + \mathbf{n}(k,t). \quad (1)$$

The vector $\mathbf{s}(k,t)$ consists of the source spectrum $S_n(k,t)$ as $\mathbf{s}(k,t) = [S_1(k,t), \cdots, S_N(k,t)]^T$. The symbol k and t denote the indices for the frequency and the time-frame, respectively. The vector $\mathbf{n}(k,t)$ denotes the background noise. The matrix $\mathbf{A}_k$ is the transfer function matrix, its (m,n) element being the transfer function of the direct path from the nth source to the mth microphone. The nth column vector of $\mathbf{A}_k$ is termed the location vector of the nth source.

2.2. Sound Localization

The spatial spectrum is obtained as an average of the MUSIC spatial spectrum as

$$\bar{P}(r,\theta) = \frac{1}{K} \sum_{k=k_L}^{k_H} P(r,\theta,k). \quad (2)$$

The symbols k_L and k_H are indices for the lower and the upper boundaries of the frequency range, and $K = k_H - k_L + 1$. The symbols r and θ denote the range and the direction, respectively. The MUSIC spatial spectrum is given by

$$P(r,\theta,k) = \frac{1}{|\tilde{\mathbf{a}}_k^H(r,\theta)\mathbf{E}_k^n|^2}. \quad (3)$$

where

$$\tilde{\mathbf{a}}_k(r,\theta) = \frac{\mathbf{a}_k(r,\theta)}{\|\mathbf{a}_k(r,\theta)\|} \quad (4)$$

is the normalized location vector for the scanning point (r,θ) [1]. The matrix $\mathbf{E}_k^n$ consists of the eigenvectors obtained by the eigenvalue decomposition of $\mathbf{R}_k$ as

$$\mathbf{R}_k = \mathbf{E}_k \mathbf{\Lambda}_k \mathbf{E}_k^{-1}, \quad (5)$$

where $\mathbf{R}_k$ is the spatial correlation matrix defined as

$$\mathbf{R}_k = E[\mathbf{x}(k,t)\mathbf{x}^H(k,t)]. \quad (6)$$

The eigenvectors are split as $\mathbf{E}_k = [\mathbf{E}_k^s | \mathbf{E}_k^n]$ where $\mathbf{E}_k^s$ and $\mathbf{E}_k^n$ denote the sets of eigenvectors corresponding to the N dominant eigenvalues and the rest of the eigenvalues, respectively. The location of sources can be known from the peak of this spatial spectrum.

2.3. Sound Separation

By using MVBF, the spectrum of the nth sound source is recovered by the following filtering:

$$\hat{S}_n(k,t) = \mathbf{w}^H(k)\mathbf{x}(k,t) \quad (7)$$

where

$$\mathbf{w} = \frac{\mathbf{R}_k^{-1}\hat{\mathbf{a}}_{n,k}}{\hat{\mathbf{a}}_{n,k}^H \mathbf{R}_k^{-1} \hat{\mathbf{a}}_{n,k}}. \quad (8)$$

The vector $\hat{\mathbf{a}}_{n,k}$ is the location vector for the nth source estimated by the sound localization. Two modified versions of MVBF were employed in this paper.

Table 1: Differences of MV1 and MV2.

	MV1	MV2
Correlation	$\mathbf{K}_k$	$\mathbf{Q}_k$
Advantage	high noise reduction	high tracking capability
Disadvantage	absence of target must be detected	performance depends on localization accuracy

Table 2: Results of the benchmark test. The processing time for 1 s data is shown.

	Time [s]	Processor
FFT & Correlation	0.86	DSP-B
MUSIC	0.34	Host
MVBF	0.16	Host
Filtering	0.52	DSP-A

When the spatial correlation matrix $\mathbf{K}_k$ is used instead of $\mathbf{R}_k$ in (8) as

$$\mathbf{w}_{MV1} = \frac{\mathbf{K}_k^{-1}\hat{\mathbf{a}}_{n,k}}{\hat{\mathbf{a}}_{n,k}^H \mathbf{K}_k^{-1}\hat{\mathbf{a}}_{n,k}}, \tag{9}$$

(8) becomes the maximum likelihood (ML) estimator [3]. The matrix $\mathbf{K}_k$ denotes the spatial correlation that is estimated in the period where the target signal $S_n(k,t)$ is absent. The ML estimator $\mathbf{w}_{MV1}$ yields a high noise reduction performance with a relatively small amount of data compared with (8) and is termed "MV1" hereafter.

On the other hand, the authors previously proposed the modified MVBF [1], in which the spatial correlation $\mathbf{R}_k$ is replaced by the following synthesized spatial correlation:

$$\mathbf{Q}_k = \hat{\mathbf{A}}_k \hat{\mathbf{A}}_k^H + \gamma \mathbf{I}. \tag{10}$$

The second modified MVBF using this $\mathbf{Q}_k$ is termed "MV2", and is given by

$$\mathbf{w}_{MV2} = \frac{\mathbf{Q}_k^{-1}\hat{\mathbf{a}}_{n,k}}{\hat{\mathbf{a}}_{n,k}^H \mathbf{Q}_k^{-1}\hat{\mathbf{a}}_{n,k}}. \tag{11}$$

As described above, a salient feature of MV2 is the use of $\mathbf{Q}_k$ that is synthesized using the estimate of the localization, $\hat{\mathbf{A}}_k$. The estimate of the localization can be obtained with less input observation (0.2-0.5s) [1] compared to that for K_k (more than 1 s). Thus, the advantage of this method is its faster capability of tracking environmental changes compared with MV1. This is a desired feature for applications such as the mobile robot. The differences of MV1 and MV2 are summarized in Table 1.

The second term in (10) corresponds to the correlation for the background noise. The parameter γ functions as weight in the weighted least-square problem. When the real background noise is high, the reduction performance for the background noise can be raised by employing a larger γ.

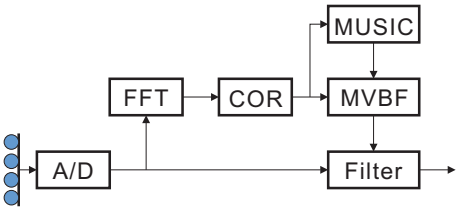


Figure 1: Block diagram of the proposed system.

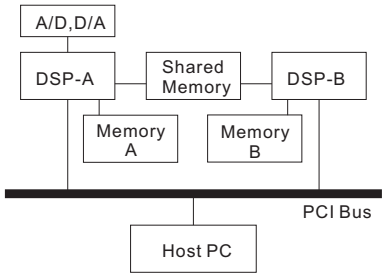


Figure 2: Architecture of the DSP system.

3. System

3.1. Hardware

A block diagram of the system is depicted in Fig. 1. First, the microphone-array input is Fourier-transformed and the spatial correlation is calculated by the FFT and COR modules. Then the spatial correlation $\mathbf{R}_k$ is sent to the MUSIC module and the source location ($\hat{\mathbf{A}}$) is estimated. In the MVBF module, the beamformer coefficients are calculated. For MV1, the correlation in the absence of the target, $\mathbf{K}_k$, is also estimated in the COR module and is sent to the MVBF module. The length of data used for estimating $\mathbf{R}_k$ and that used for $\mathbf{K}_k$ is 1.0 s and 0.5 s, respectively, in this paper. The estimated coefficients are transformed to the time domain by Fourier transform and the input signal is filtered in Filter with these time-domain coefficients.

The system consists of dual DSPs (TI C6701) and the host PC (Pentium III 600MHz). The architecture of the system is depicted in Fig. 2. The assignment of the computational resources and the results of the benchmark tests are shown in Table 2.

The microphone array was circular in shape with a diameter of 0.5 m and $M = 8$ and was mounted on a mobile robot XR-4000 as depicted in Fig. 4(b). The frequency range of averaging in (2) was [500, 3000] Hz.

3.2. Location Vector Database

Prior to the operation of the system, the location vectors for the possible source locations were measured as a database for the source localization and separation. The data were taken in the range of 60-160 cm as indicated in Table 3. The total number of the data entries is 492 points. The database is used for $\mathbf{a}_k(r,\theta)$ in (4) for the sound localization, and $\hat{\mathbf{a}}_{n,k}$ in (9) and $\hat{\mathbf{A}}_k$ in (10) for the sound separation. The measured data and the details of the measurement are available from the RWCP sound scene database web-site [4].

Table 3: Location vector database.

	Direction	Range
60-90 cm range	every 10 °	every 10 cm
100-160 cm range	every 5 °	every 20 cm

Table 4: Source Location

	Source 1		Source 2	
	Angle	Range	Angle	Range
Real	10	80	70	120
Estimated	10	90	70	140

3.3. Online ASR System

The above sound source localization/separation system is connected through a network to the ASR server currently being developed by the authors as depicted Fig. 3. The ASR system, denoted as RVCP-niNja hereafter, consists of a discrete-HMM ASR engine (niNja) [5] and a network interface with the network protocol specially designed for exchanging speech information through the network (Remote Voice Control Protocol, RVCP) [6]. This ASR system is useful for applications such as the mobile robot employed in this paper which have limited local resources. By using RVCP, the ASR system itself can be distributed over several computational resources on a network, a useful feature for the cases with heavy ASR load such as multiple mobile robots.

4. Evaluation

4.1. Condition

The evaluation experiment was conducted in an ordinary office. The configuration of the source and the microphone array is shown in Fig. 4. The location of the sound sources (loudspeakers) is shown in Table 4. The sound sources, S1 and S2, emit speech and music respectively. The sound levels of S1, S2 and the background noise at the position of microphone #1 were 68, 67, 48 [dBA], respectively. The background noise mainly consists of the noise of the PC fan.

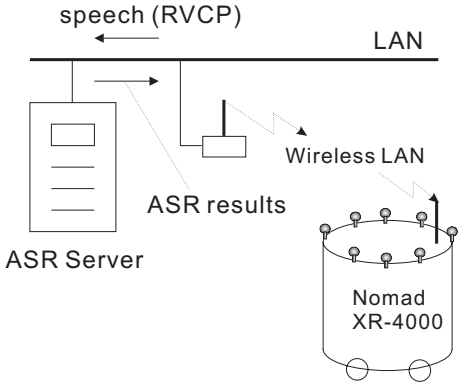
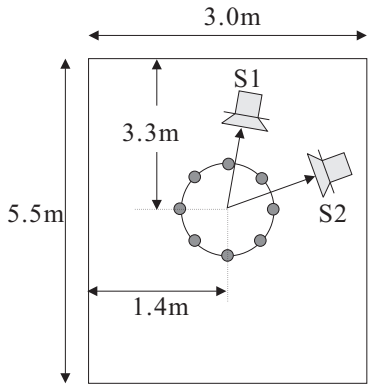


Figure 3: Online ASR system.



(a) Configuration of sound sources

(b) Scene of experiment

Figure 4: Experimental setup.

4.2. Results

Figure 5 shows the obtained spatial spectrum. From the peak position of this spectrum, the location of the sources was estimated. The estimated location is shown in Table 4. It can be seen that the estimation of the direction is accurate while the range was estimated as being a little larger. This is considered to be mainly an effect of room reflections.

Figure 6 shows the recognition rate for 492 Japanese words. For the sake of comparison, an off-line continuous-HMM speech recognizer (HTK [7] with IPA phonetic model [8]) was employed as well as RVCP-niNja. When MV1 was employed, a high recognition rate was achieved. On the other hand, when employing MV2, the recognition rate was reduced by around 20%. As described in Table 1, this is due to the accuracy of the sound localization. As indicated in Table 3, the location vectors were measured every 5° in the range of 1-1.6 m. This means that, even when the localization is the most accurate, the estimated location and the real location could differ by ± 2.5° .

Figure 7 shows the gain of MV2 in the vicinity of the interference direction. From this figure, it can be seen that, when the real and the estimated location differ by ± 2.5° , the noise reduction performance is reduced to around 10 dB at the higher frequencies. When higher performance is required, a database with higher resolu-

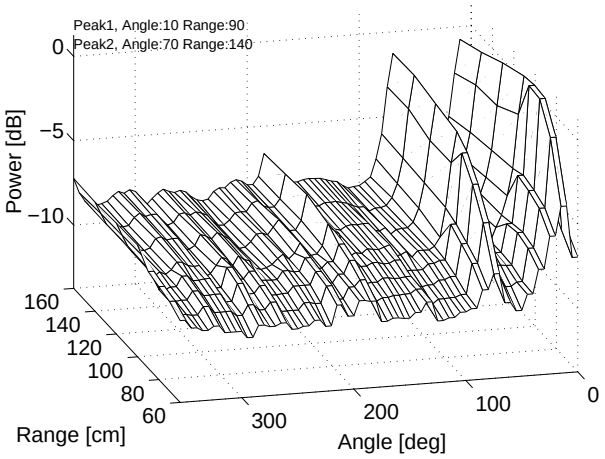


Figure 5: Spatial spectrum obtained by 2D-MUSIC.

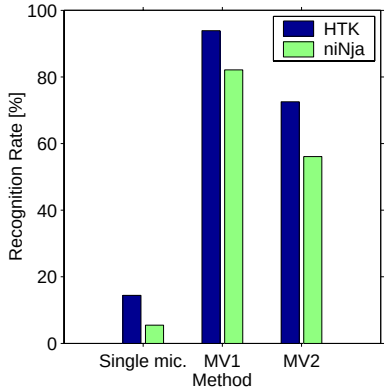


Figure 6: ASR rate.

tion (e.g., every 2.5°) should be employed.

5. Discussion and Conclusion

In this paper, a real-time system of sound source localization/separation for sources in the near-field range was developed and evaluated. The system consists of dual DSPs and a PC.

Regarding the sound localization, the directions of the two sound sources were accurately estimated. As for the range estimation, the resolution was lower compared with that of the direction estimation, due to the variation of phase difference between the microphones being smaller. The accuracy of the range estimation is even lower than those of the off-line experiments in [1]. This is considered to be due mainly to the higher background noise and the early reflections as compared with the environment used in [1].

Regarding the sound separation, a high recognition rate was achieved when MV1 was employed. Future work on MV1 includes the detection of the presence/absence of the target signal. MV1 requires both $\mathbf{R}_k$ (spatial correlation for the target being present) for determining the beam (focus) location and $\mathbf{K}_k$ (spatial correlation for the target being absent) for determining the null location. In the present study, the presence/absence of the target signal is indicated to the system by a human operator.

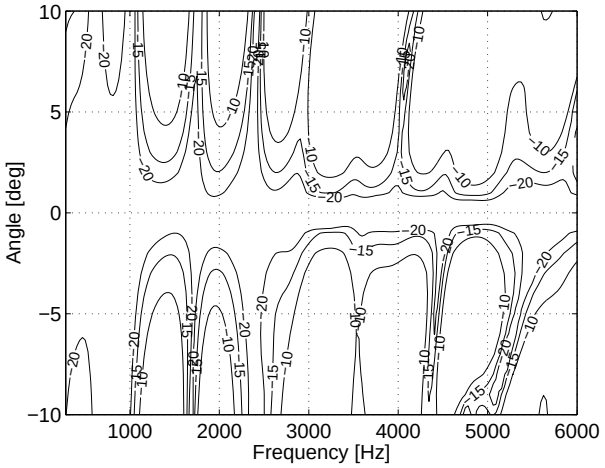


Figure 7: Gain of MV2 in the vicinity (±10°) of the interference direction.

Automation of this detection is required.

When employing MV2 for the sound separation, the performance was reduced by around 20%. This is a trade-off between high noise reduction performance and high tracking capability. MV2 requires only $\mathbf{R}_k$. In the off-line experiment, it was shown that the beamformer coefficients were able to be updated every 0.2 s [1]. However, the evaluation of the tracking capability such as that for moving sources has not yet been conducted and remains for future study.

6. References

[1] Futoshi Asano, Hideki Asoh, and Toshihiro Matsui, "Sound source localization and separation in near field," IEICE Trans. Fundamentals, vol. E83-A, no. 11, pp. 2286–2294, November 2000.

[2] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," IEEE Trans. Antennas Propag, vol. AP-34, no. 3, pp. 276–280, March 1986.

[3] Don H. Johnson and Dan E. Dudgeon, Array signal processing, Prentice Hall, Englewood Cliffs NJ, 1993.

[4] http://tosa.mri.co.jp/sounddb/.

[5] K. Itou, S. Hayamizu, K. Tanaka, and H. Tanaka, "System design, data collection and evaluation of a speech dialog system," IEICE Trans. INF & SYST., vol. E76-D, no. 1, pp. 121–127, Jan. 1993.

[6] Masataka Goto, Katunobu Itou, Tomoyosi Akiba, and Satoru Hayamizu, "Speech completion: New speech interface with on-demand completion assistance," in Proc. of HCI International 2001, 2001.

[7] http://htk.eng.cam.ac.uk/.

[8] T. Kawahara, T. Kobayashi, K Takeda, N. Minematsu, K. Itou, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano, "Japanese dictation toolkit: Plug-and-play framework for speech recognition r&d," in Proc. of ICASSP'99, March 1999, pp. I–393.