Acoust. Sci. & Tech. **30**, 5 (2009)

ACOUSTICAL LETTER

Improving the robustness of multiple signal classification (MUSIC) method to reflected sounds by sub-band peak-hold processing

Takashi Suzuki* and Yutaka Kaneda†

Department of Information and Communication Engineering, Graduate School of Eng, Tokyo Denki University, 2–2 Kanda-Nishiki-cho, Chiyoda-ku, Tokyo, 101–8457 Japan

(*Received 24 March 2009, Accepted for publication 1 May 2009*)

Keywords: Reflected sound, DOA estimation, MUSIC, Microphone array, Sub-band peak hold
PACS number: 43.60.–c, 43.60.Jn, 43.60.Gk [doi:10.1250/ast.30.387]

1. Introduction

In estimating the direction of a sound source in an ordinary room, the accuracy of the estimate deteriorates because of the effect of reflected sounds. To solve this problem, we proposed sub-band peak-hold (SBPH) processing and applied it to the correlation method using two microphones, thereby verifying its validity [1,2]. In this article, we report the successful application of SBPH processing to the multiple signal classification (MUSIC) method using multiple microphones.

2. Estimation of sound source direction by MUSIC [3]

First, we calculate the cross spectrum $\Phi_{ij}(\omega)$ ($i, j = 1, 2, \ldots, M$; M is the number of microphones) for the i-th and j-th microphone outputs, $X_i(\omega)$ and $X_j(\omega)$, as

$$\Phi_{ij}(\omega) = E[X_i^*(\omega)X_j(\omega)]. \tag{1}$$

Here, * and $E[\cdot]$ represent the conjugate complex and expected value, respectively.

Then, we introduce a spatial correlation matrix $\boldsymbol{R}(\omega)$ consisting of $\Phi_{ij}(\omega)$ as its ij-element. Using the orthogonal property between the signal subspace and noise subspace of the spatial correlation matrix $\boldsymbol{R}(\omega)$, we obtain the spatial spectrum (i.e. estimated result of the directions of sound sources) at frequency ω, $P(\omega, \theta)$, as

$$P(\omega, \theta) = 1/\{\boldsymbol{d}^{\mathrm{H}}(\omega, \theta)\boldsymbol{R}_N(\omega)\boldsymbol{d}(\omega, \theta)\}, \tag{2}$$

$$\boldsymbol{R}_N(\omega) = \sum_{q=L+1}^{M} \boldsymbol{v}_q(\omega)\boldsymbol{v}_q{}^{\mathrm{H}}(\omega), \tag{3}$$

where H represents the conjugate transpose and $\boldsymbol{v}_q(\omega)$ ($q = L+1, \ldots, M$; L is the number of sound sources) is an eigenvector for the noise subspace. $\boldsymbol{d}(\omega, \theta)$ is a steering vector defined by

$$\boldsymbol{d}(\omega, \theta) = [1, e^{-j\omega\tau_2(\theta)}, \ldots, e^{-j\omega\tau_M(\theta)}]^{\mathrm{T}}. \tag{4}$$

Here, T represents the transpose, and $\tau_i(\theta)$ shows the signal delay time (relative value) at the i-th microphone for sound arriving from direction θ.

In the case of a broadband sound source, the spatial spectrum $P(\theta)$ showing the direction of the sound source can be obtained by adding $P(\omega, \theta)$ in the designated frequency range of ω_1–ω_2, as shown in Eq. (5).

$$P(\theta) = \sum_{\omega=\omega_1}^{\omega_2} P(\omega, \theta) \tag{5}$$

3. SBPH processing [1,2]

Figure 1(a) shows an example of a received sound waveform $x_1(t)$ that consists of a pulsive direct sound and a reflected sound. Figure 1(b) shows the waveform obtained after peak-hold processing of the received sound. As shown, peak-hold processing masks the reflected sound with small amplitude, that follows direct sound, by maintaining the direct sound's amplitude.

Furthermore, the effect of reflected sounds with large amplitudes is reduced by taking the logarithm of the peak amplitude after peak-hold processing (a logarithmic operation). Then, the low-frequency components of peak-hold signals are suppressed by time difference. To utilize the feature of speech signal that is its frequency components arrive at different times for different frequency bands, above mentioned peak-hold processing is applied to sub-band signals.

4. Proposed method (SBPH-MUSIC)

Figure 2 shows a block diagram of the signal processing used for the MUSIC method to which SBPH is applied. First, we applied the short-time Fourier transform (STFT) to received signals, and output the time series of the amplitude components of sub-band signal $|X_i(k, t)|$ (i: microphone number, k: sub-band number, t: discrete time). Subsequently, peak-hold processing (PH), logarithmic operation (log), and time difference (Diff) are applied to generate signal $|X_i(k, t)|_p$, in which the effect of reflected sound is reduced. Here, sub-band signal $|X_i(k, t)|_p$ in each channel show narrow-band time-series waveform.

Next, for the k-th sub-band (the sub-band number is fixed), we obtain the cross spectrum $\Phi_{ij}(k, \omega')$ between the time-series waveform for the i-th channel $|X_i(k, t)|_p$ and that for the j-th channel $|X_j(k, t)|_p$ by

$$\Phi_{ij}(k, \omega') = E[F[|X_i(k, t)|_p]^* \cdot F[|X_j(k, t)|_p]], \tag{6}$$

where $F[\cdot]$ and ω' represent the short-time Fourier transform and the frequency of signal $|X_i(k, t)|_p$, respectively.

*Current affiliation: The Tokyo Electric Power Company, Inc.
†e-mail: kaneda@c.dendai.ac.jp

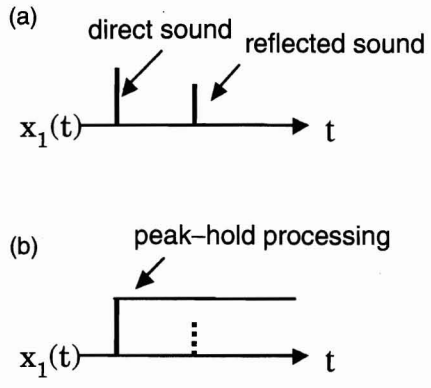

Fig. 1 Peak-hold processing.

This operation is performed for all sub-bands, and the total sum for k is designated as $\Phi_{ij}(\omega')$.

$$\Phi_{ij}(\omega') = \sum_{k} \Phi_{ij}(k, \omega') \tag{7}$$

Then, we represent the spatial correlation matrix consisting of $\Phi_{ij}(\omega')$ as its ij-element at frequency ω' by $\boldsymbol{R}(\omega')$. Similarly to Eqs. (2)–(5) in Section 2, we calculate the spectrum of the sound source direction $P(\theta)$ using $\boldsymbol{d}(\omega', \theta)$ and $\boldsymbol{R}_N(\omega')$ obtained from $\boldsymbol{R}(\omega')$.

Table 1 Experimental condition.

Room dimensions	5.0 [W] × 6.0 [D] × 2.5 [H] [m]
Reverberation time	0.38 s
SNR	25 ~ 30 dB
Array height	1.2 m
Number of microphones	8
Array shape	circular
Diameter of array	$d = 0.3$ m
Distance of sound source	$r = 3.0$ m
Direction of sound source	$\theta_s = 0, 30, 45$ deg.
Sampling frequency	32 kHz
Frame Length for STFT	$T = 32$ [sample]
Shift length for STFT	$T/8$
Frequency range of MUSIC method	$\omega = 500 \sim 3{,}000$ Hz
Frequency range of SBPH-MUSIC method	$\omega' = 500 \sim 2{,}000$ Hz
Assumed number of sound sources	$L = 1$

5. Experiment in actual environment

Table 1 shows the experimental conditions adopted. Note, in the proposed method, MUSIC is applied to sub-band amplitude signal $|X_j(k,t)|_p$, whose frequency component covers up to about 2,000 Hz. And the frequency bands for MUSIC and SBPH-MUSIC shown in Table 1 are decided experimentally so as to derive the best results.

The first experiment was carried out by placing a microphone array at the center of the room. Then, as the second experiment, the array was placed in a corner of the room, where the effect of early reflected sounds is large (Fig. 3).

As a sound source, a person (male) uttered 30 words, three times for each word and angle (three directions); the total number of measurements was 270. (In the first experiment, for 30 words in three directions, the total number of utterances was 90.) The performances of conventional MUSIC and SBPH-MUSIC methods were compared under these conditions.

Figure 4 shows the percentage of correct estimations in the experiment (error margin: 10 deg). Figures 4(a) and 4(b) show the results of the first and second experiments, respectively. In the first experiment in which the effect of early reflected sound is small, both the conventional

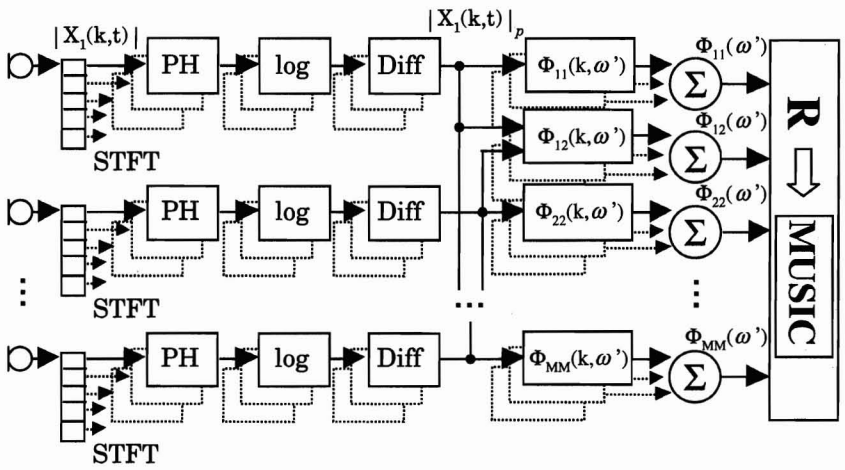

Fig. 2 Block diagram of the proposed method.

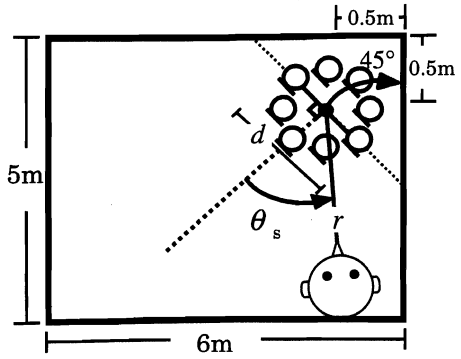

Fig. 3 Microphone arrangement for the second experiment.

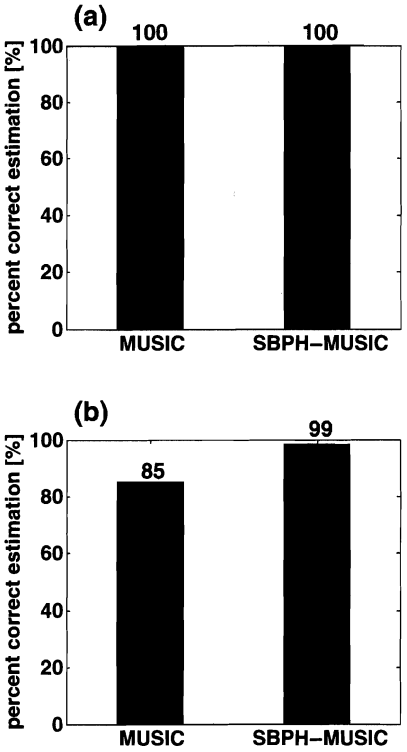

Fig. 4 Correct estimation rate of DOA experiment. (a) Experiment 1, (b) Experiment 2.

(MUSIC) and proposed (SBPH-MUSIC) methods yielded 100% correct estimation. In contrast, in the second experiment, in which the effect of the early reflected sound is large, the percentage of correct estimations dropped to 85% with the conventional method. However, it remained at approximately 99% with the proposed method (here, the percentage of correct estimations for a 5 deg error margin was 96%), demonstrating an improvement of robustness against reflected sounds.

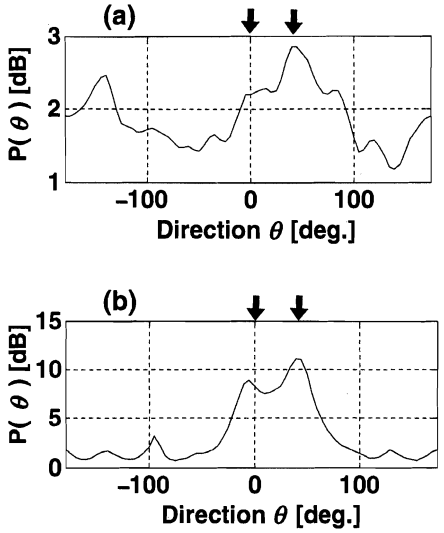

Fig. 5 Spatial spectra $P(\theta)$. (a) MUSIC, (b) SBPH-MUSIC.

6. Comparison of spatial spectra for multiple sound sources

We compared spatial spectra when there were two sound sources and the microphone array was placed near a wall. In this experiment, the distances from each sound source were both 3 m away from the array and the directions of the sound sources were 0 and 45 deg. Assumed number of sound sources for MUSIC, L, was set to 2. Figures 5(a) and 5(b) show the spatial spectra for the MUSIC and SBPH-MUSIC methods, respectively. The black arrows above these figures represent the directions of the sound sources.

According to Fig. 5, the highest peak appears in the 45 deg direction in the MUSIC method; however, incorrect peaks are produced owing to the effect of reflected sounds (e.g., in the direction of −140 deg). On the other hand, in the SBPH-MUSIC method, although minor directional error is observed owing to the effects of reflected sounds, two clear peaks are observed in the directions of the sound sources.

7. Summary

In this study, we applied SBPH processing, which is robust against reflected sounds, to the MUSIC method. The results show that it is possible to maintain high-level performance in estimating the sound source direction, compared with the conventional MUSIC method, even in room environments in which the effect of early reflected sound is significant.

References

[1] Y. Kaneda, "Sound source localization for wide-band signals under a reverberant condition," *J. Acoust. Soc. Jpn. (E)*, **14**, 47–48 (1993).

[2] T. Suzuki and Y. Kaneda, "A study of DOA estimation based on sub-band peak hold processing," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 751–752 (2007).

[3] M. Brandstein and D. Ward, Eds., *Microphone Arrays* (Springer-Verlag, New York, 2001), p. 186.