https://doi.org/10.20965/jrm.2017.p0072

Paper:

Influence of Different Impulse Response Measurement Signals on MUSIC-Based Sound Source Localization

Takuya Suzuki*, **Hiroaki Otsuka***, **Wataru Akahori***, **Yoshiaki Bando****, **and Hiroshi G. Okuno*****

*Faculty of Science and Engineering, Waseda University
3-4-1 Okubo, Shinjuku, Tokyo 169-8555, Japan
E-mail: {takuya-suzuki.ph@toki, akahori@akane}.waseda.jp, ootsuka_w@icloud.com
**Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto 606-8501, Japan
E-mail: yoshiaki@kuis.kyoto-u.ac.jp
***Graduate Program for Embodiment Informatics, Waseda University
2-4-12 Okubo, Shinjuku, Tokyo 169-0072, Japan
E-mail: okuno@nue.org

[Received August 1, 2016; accepted October 27, 2016]

Two major functions, sound source localization and sound source separation, provided by robot audition open source software HARK exploit the acoustic transfer functions of a microphone array to improve the performance. The acoustic transfer functions are calculated from the measured acoustic impulse response. In the measurement, special signals such as Time Stretched Pulse (TSP) are used to improve the signal-to-noise ratio of the measurement signals. Recent studies have identified the importance of selecting a measurement signal according to the applications. In this paper, we investigate how six measurement signals – up-TSP, down-TSP, M-Series, Log-SS, NW-SS, and MN-SS – influence the performance of the MUSIC-based sound source localization provided by HARK. Experiments with simulated sounds, up to three simultaneous sound sources, demonstrate no significant difference among the six measurement signals in the MUSIC-based sound source localization.

Keywords: robot audition, impulse response measurement techniques, acoustic transfer function, sound source localization, multiple signal classification (MUSIC)

1. Introduction

"Robot Audition" recognizes a mixture of sounds captured by a set of synchronized microphones (hereinafter, *a microphone array*) to understand auditory environments. This capability is crucial for symbiosis between a service robot and people [1, 2]. During interaction with people, a robot hears his/her utterance in addition to environmental sounds; at times, it encounters cases where people interrupt the robot's utterance. With microphones on a robot rather than attached to each person, robot audition can enable more flexible interactions with people. In multiparty interactions, robot audition enables the ability to listen to several things simultaneously, whereas conventional systems frequently focus on a particular person, discarding other utterances.

Nishimuta et al., for example, developed an interactive quizmaster robot system called "HATTACK25" that can manage a multiparty speech-based quiz game similar to a Japanese TV program "ATTACK25" and an American program called "Jeopardy!" [3]. HATTACK25 provides two modes, school-class-type and auction-type. In the former, participants first attempt to acquire the right to answer by saying "yes." In the latter, they can answer a question directly. In either case, the quizemaster robot with robot audition localizes the respondents and recognizes who said what first.

Robots, like people, hear a mixture of sounds in their daily lives. Therefore, robot audition should provide three fundamental functions: sound source localization (*SSL*), sound source separation (*SSS*), and automatic speech recognition (*ASR*) [1, 3, 4]. Some open source robot audition software provides only SSL; other provides SSS [4]. However, only HARK (Honda Research Institute Japan Audition for Robots with Kyoto University) [5] supports all three functions. The SSL and SSS provided by HARK exploit acoustic transfer functions (hereinafter, *transfer functions*) to improve the performance [6]. In fact, HARK SSL is made robust by implementing MUSIC (Multiple Signal Classification) [7] with the transfer functions [8].

A transfer function specifies the spectral characteristics for how a signal transfers from a sound source to a microphone. It is calculated from an impulse response measured using an impulse response measurement signal (hereinafter, *measurement signal*) [9].

The impulse response is crucial in both analyzing acoustic fields and synthesizing binaural sounds and acoustic fields [10, 11]. In virtual reality, three-dimensional (3D) acoustic field synthesis improves the reality of 3D images. Because the quality of analysis and synthesis depends on the impulse response, measurement signals have recently received more attention. Kaneda stated [12]:

> "Because the error that can be tolerated depends on the applications of the impulse response, guidelines for the selection of measurement signals for different applications should be established."

In this paper, we investigate the influence of measurement signals on the performance of MUSIC-based SSL. Nakamura et al. confirmed that transfer functions calculated from coarsely measured impulse responses with linear interpolation in temporal- and frequency-domains with hierarchical SSL work effectively for SSL and SSS in HARK [13]. However, the influence of measurement signals has not been evaluated. The remainder of this paper is organized as follows: Section 2 describes the background and related work; Section 3 presents the measurement of impulse responses; Section 4 presents the experimental evaluation; and Section 5 concludes the paper.

2. Background and Related Work

This section describes background knowledge including impulse responses and the MUSIC-based SSL of HARK.

2.1. Impulse Response Measurement Signals

A transfer function is obtained directly from an impulse response. When an input $x(t)$ is provided to the system, it generates an output $y(t)$ where its impulse response is $h(t)$.

$$y(t) = \int_{-\infty}^{\infty} h(\tau)x(t-\tau)d\tau. \quad \dots\dots\dots \quad (1)$$

By transforming Eq. (1) to the frequency domain using the Fourier transform, the following equation is obtained.

$$Y(j\omega) = H(j\omega)X(j\omega), \quad \dots\dots\dots \quad (2)$$

where j is an imaginary unit, ω is the frequency, and $H(j\omega)$ and $X(j\omega)$ are the Fourier transforms of $h(\tau)$ and $x(\tau)$, respectively. Eq. (2) implies that if $x(t)$ is a δ-function, the transfer function is the impulse response itself.

Because it is difficult to generate an ideal impulse signal, acoustic communities have studied how to improve the quality of an impulse response, for example, the signal-to-noise ratio (SNR) of the measured signals. The two main approaches use either pseudo-random white noise or time varying frequency signals [14]. Schroeder proposed an M-Series signal based on pseudo-random noise in 1979 [15] and then proposed a less computational technique [16]. To overcome the unrepeatability of M-Series signals due to pseudo-random white noise, Aoshima proposed a time-varying frequency signal called "Time-Stretched Pulse" (*TSP*) to provide the repeatability of measurements [17]. Then, Suzuki et al. extended his idea to propose a TSP suitable for measuring in large halls [18]. Another well-known variation based on time-varying frequency signals is SineSweep [19].

Recently, Farina proposed logarithmic SineSweep (*Log-SS*) to overcome the distortion artifacts of the other techniques that appear in the deconvoluted impulse responses when linear and time-invariant assumptions do not hold [20].

Stan et al. compared four different impulse response-measuring techniques, Maximum Length Sequence, Inverse Repeated Sequence, Time-Stretched Pulses, and SineSweep [14]. The first two techniques are based on pseudo-random white noise. They concluded that the first two techniques appear to be more accurate in the presence of non-white noise and that the latter two seem most appropriate in a quiet environment.

2.2. Six Impulse Response Measurement Signals

We choose the following six measurement signals based on the previous discussion and Kaneda's paper [12]:

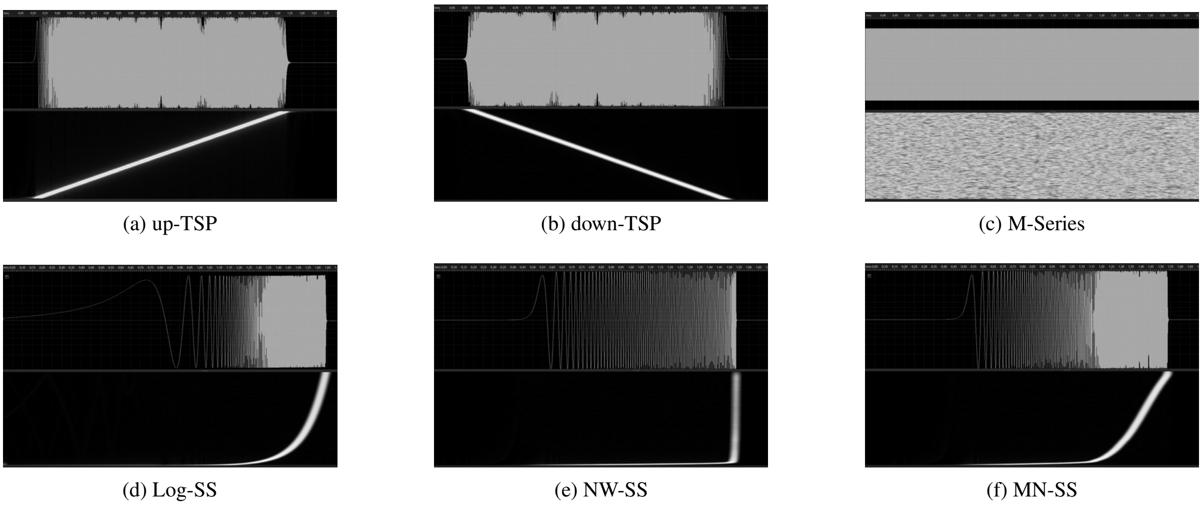
- upward TSP (*up-TSP*),
- downward TSP (*down-TSP*),
- M Series signals (*M-Series*),
- Logarithmic SineSweep (*Log-SS*, or Pink-TSP),
- Noise Whitening SineSweep (*NW-SS*) [10], and
- Minimum Noise SineSweep (*MN-SS*) [21].

The spectrogram of each signal is presented in **Fig. 1**. Two new variants of SineSweep, NW-SS and MN-SS, are introduced to address non-stable noise at the time of measurement by whitening ambient noise and by minimizing ambient noise, respectively (see **Figs. 3(a)** and **(b)** in Section 3.2). In general, up-TSP and down-TSP may have harmonic distortion in the negative and positive time axis, respectively. Log-SS improves a signal-to-noise ratio in lower frequencies and may suppress harmonic distortion. M-Series may have non-linear distortion.

2.3. HARK Sound Source Localization

HARK provides a MUSIC-based algorithm for SSL [7, 8]. The original MUSIC assumes the following: (1) the number of sound sources and noise are fixed in advance, (2) the power of any sound source exceeds that of the noise, and (3) the number of sound sources does not exceed that of the microphones. Here, noise means only directional noise and does not include ambient noise. Because MUSIC-based SSL outperforms conventional beamforming [22] if the above assumptions hold, HARK recommends using the MUSIC-based SSL [23]. Further, it does not assume (3) above. HARK provides precalculated transfer functions for various microphone arrays measured by up-TSP.

First, eigenvalues and eigenvectors are calculated by decomposing an input correlation matrix. It is represented as $[\boldsymbol{e}_i(\omega), \dots, \boldsymbol{e}_M(\omega)]$ with eigenvalues $\lambda_1, \dots, \lambda_M$, where


(a) up-TSP (b) down-TSP (c) M-Series

(d) Log-SS (e) NW-SS (f) MN-SS

Fig. 1. Wave form and spectrogram of six measurement signals for first 1.65 s.

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M$. Let M and L be the number of microphones and number of sound sources, respectively. Because the corresponding eigenvalue represents the power of each eigenvector, the above assumptions specify that λ_i for $1 \leq i \leq L$ represents the eigenvalue of the i-th sound source and λ_i for $L+1 \leq i \leq M$ represents that of the noise.

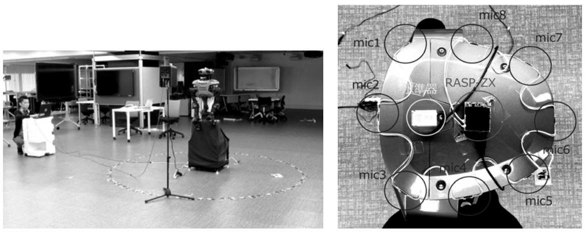
The spatial spectrum (hereinafter, *MUSIC spectrum*) $P(\theta)$ of the direction θ is defined as follows:

$$P(\theta) = \sum_{\omega} \frac{|H^{\mathrm{H}}(\theta,\omega)H(\theta,\omega)|}{\sum_{i=L+1}^{M} |H(\theta,\omega)^{\mathrm{H}} \boldsymbol{e}_i(\omega)|} \quad . \; . \; . \; . \; . \; . \quad (3)$$

where H represents the conjugate transpose operator, $H(\theta,\omega)$ is a transfer function of θ and ω, and $\boldsymbol{e}_i(\omega)$ is an eigenvector of the input correlation matrix. $H(\theta,\omega)$ is also referred to as a *steering vector* towards the direction θ at the frequency of ω. When the direction of the steering vector $H(\theta,\omega)$ matches that of a sound source, the MUSIC spectrum $P(\theta)$ becomes infinity; this is an ideal case. Typically, ambient noise is not white noise and thus noise is cross-correlated with the sound sources. Therefore, the denominator of Eq. (3) does not become zero.

HARK uses 32 points in a 16 kHz sampling of an impulse response to obtain the transfer functions for SSL. It uses 512 points of an impulse response for SSS. These two kinds of transfer function are generated from impulse responses using HARK tools [5].

HARK provides various parameters to facilitate the control of MUSIC-based SSL [a]. Important parameters of the module *LocalizeMUSIC* include *NUM_SOURCE* to specify the number of sound sources (default: 2) and *LOWER_BOUND_FREQUENCY* and *UPPER_BOUND_FREQUENCY* to specify the lower and upper bound frequency in the signal processing (default: 500 Hz and 2800 Hz, respectively). Important parameters of *SouceTracker* are *THRESHOLD* to indicate that a sound source is ignored if its MUSIC spectrum is less than this value, *PAUSE_LENGTH* to specify the lifetime of a sound source, and *MIN_SRC_INTERVAL* to specify the threshold value of angular difference for determining if the sound source is the same as another.



(a) Snapshot of measurement (b) Microphone configuration of HIRO

Fig. 2. Recording of impulse responses.

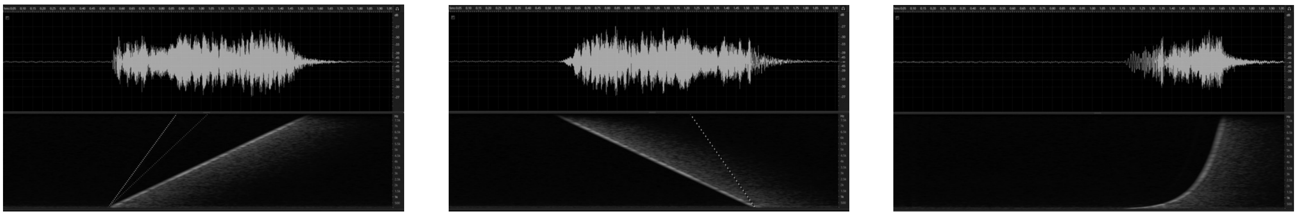
3. Measurement of Impulse Responses

This section explains the measurement of impulse responses by using an eight-element microphone array mounted on the head of a Kawada HIRO upper-torso robot.

3.1. Setting for Measurement

We used a large hall, approximately 800 m^2, to measure the impulse responses (see **Fig. 2(a)**). The HIRO was placed in the center of an approximately 5 m diameter free space. A loudspeaker, Genelec 1029A, was placed at the height of 1.5 m, 1.5 m away from the HIRO (see in **Fig. 2(a)**). **Fig. 2(b)** depicts the configuration of eight MEMS digital microphones. The eight microphones are connected to a RASP-ZX[1] that transferred eight-channel 24-bit data to a host computer by USB 2.0.

[1] http://www.sifi.co.jp/system/modules/pico/index.php?content_id=36 [Accessed October 3, 2016]

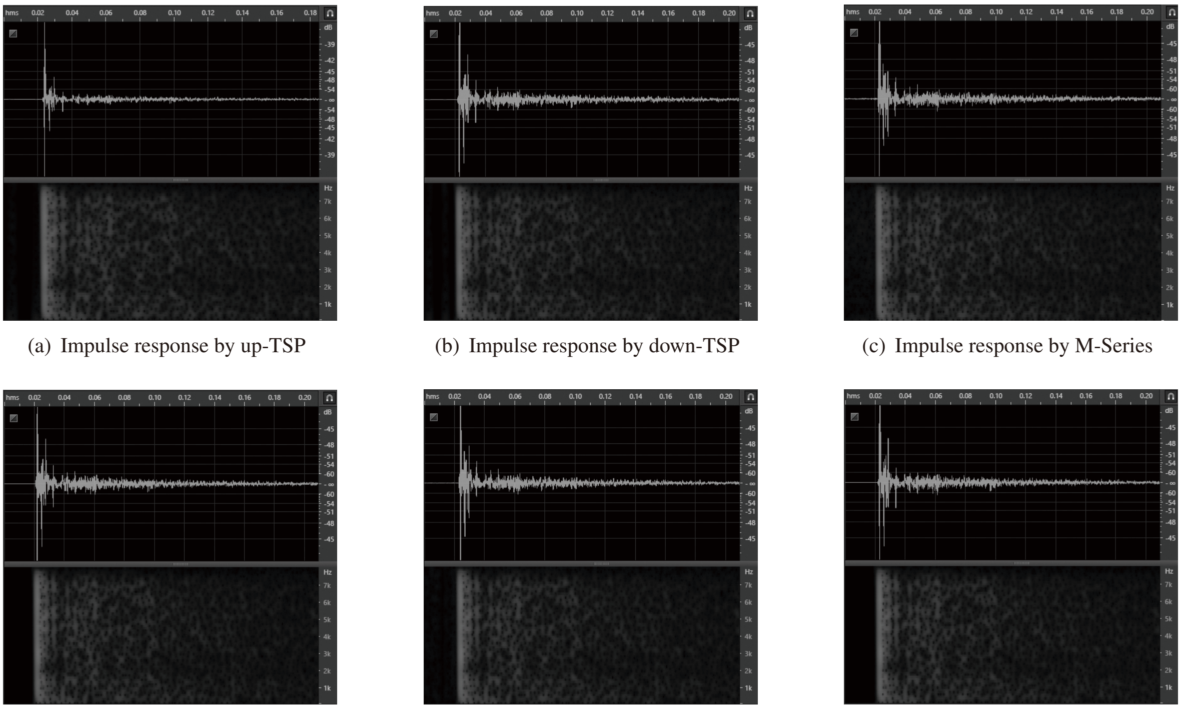


(a) Measured up-TSP Signals

(b) Measured down-TSP Signals

(c) Measured Log-SS Signals

Fig. 3. Measured signals of three measurement signals for first 1.95 s. White lines in (a) and (b) indicate harmonic distortion.



(a) Impulse response by up-TSP

(b) Impulse response by down-TSP

(c) Impulse response by M-Series

(d) Impulse response by Log-SS

(e) Impulse response by NW-SS

(f) Impulse response by MN-SS

Fig. 4. Observed impulse responses for the six measurement signals for first 2.05 s.

3.2. Measurements of Impulse Response

Benchmark signals were composed by cyclically shifting each measurement signal ten times with a 10 ms gap of smoothed offset and onset. For adaptive measurement signals, that is, NW-SS and MN-SS, ambient noise was recorded for 2 s and its mean power spectrum was obtained by averaging for 1 s [24].

Each benchmark signal was replayed from the loudspeaker every 5° around the robot and captured by the RASP-ZX as a 24-bit wave file of 16 kHz sampling. Then, a wave file was sent to a PC via USB. A multi-channel recording was made using the WIOS of HARK by playing each measuring signal. A set of impulse responses were obtained by HARKTool.[2] HARKTool first determines the channel that receives the measuring signal the earliest and thus can estimate the time of arrival at the microphone. Then, impulse responses are calculated mathematically and truncated by the time of arrival.

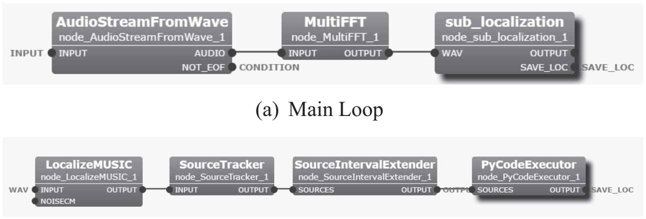
Measured signals of three measuring signals are depicted in **Figs. 3(a)–(c)**, and observed impulse responses are depicted for each measurement signal in **Figs. 4(a)–(f)**. Because obscure harmonic distortion is observed in **Figs. 3(a)** and **(b)**, white lines are added for legibility. The measured impulse response of up-TSP has second and third harmonic distortion from the beginning of the signal, whereas third harmonic distortion converges to the end of the signal. This kind of harmonic distortion may influence the performance of SSL and SSS. Because the MUSIC-based SSL of HARK uses only 32 points, that is, data of 1.953 ms, the harmonic distortion can degrade the performance of SSL.

3.3. Mixture of Sounds for Benchmark

In this subsection, the design of the benchmark sounds is described. We used one to three sound sources; their positions were as follows:

1. **One sound source**: one sound source moved from 0° to 355° by 5°. The total number of positions was 72.

[2] For WIOS and HARKTool, see HARK manual and cookbook available at http://www.hark.jp/ [Accessed October 3, 2016]



(a) Main Loop

(b) Localization with LocalizeMUSIC and SourceTracker

Fig. 5. Network for SSL by HARK.

2. **Two sound sources**: the first sound source moved in the same manner as above; the second moved from the next 5° position to the 72nd position by 5°. The total number of combinations was 5,112 $(= 72 \times 71)$.
3. **Three sound sources**: In addition to the above case, the third sound source moved from the 5° position next to the second source to the 72nd position by 5°. The total number of combinations was 357,840 $(= 72 \times 71 \times 70)$.

Speech signals were six phonetically balanced sentences extracted from ASJ-JNAS [25], each of which was spoken by three men and three women. The SNR of the speech signals to white noise are none, -10 dB, -5 dB, 0 dB, $+5$ dB, and $+10$ dB from less noisy to more noisy. Because six measurement signals and six noise ratios were examined, the number of wave files for one sound source, two sound sources, and three sound sources were 2,599, 184,032, and 12,882,240, respectively,

4. Evaluation

4.1. MUSIC Spectrums for Benchmark Sounds

Localization was conducted using the HARK network illustrated in **Fig. 5**. It receives an audio signal from an audio wave file and conducts LocalizeMUSIC with the set of parameters discussed in Subsection 4.2.

Figures 6(a)–(c) depict the MUSIC spectra with transfer functions obtained by up-TSP for one, two, and three sound sources, respectively. The SNR includes six cases: without noise (hereinafter, $+\infty$) and with -10 dB, -5 dB, 0 dB, 5 dB, and 10 dB of white noise. For a single sound source, the MUSIC spectra vary drastically owing to the influence of the power of the white noise. Conversely, they are rather stable for two and three sound sources, indifferent to the power of the white noise. This result by up-TSP is common with other measurement signals. The observation that the MUSIC spectra are stable for a mixture of sounds suggests that MUSIC-based SSL is robust against a mixture of sounds, even in noisy environments.

4.2. Criteria of Evaluation

The set of the above mixture of sounds for the benchmark were first localized by the MUSIC-based SSL of HARK. The value of *THRESHOLD* was determined empirically by verifying the localization results of all the data used in this paper: although the best value depends on each measurement signal, *THRESHOLD* was set to 28.5 for all the experiments. We used the default values for the other parameters. For example, *PAUSE_LENGTH* was set to 800 in 10-frame and *MIN_SRC_INTERVAL* 10°.

For all experiments, these values were fixed for each measurement signal. The value of *NUM_SOURCE* was fixed to three. The results of the SSL were then analyzed based on event, not frame-wise, by the following criteria proposed by Takahashi [26]:

- N: Correct,
- E5: If an error was within ±5°,
- E10: If an error was between |5°| and |10°|,
- E15: If an error was between |10°| and |15°|,
- E15-30: If an error was between |15°| and |30°|,
- I: Insertion, if an error was greater than |30°|,
- D: Deletion,
- Suffix M: In case of missing leading part.

The onset of each utterance was detected using the corresponding sound source separated by HARK. If the onset was not the same as the original signal, the suffix M was added. Note that N does not contain NM. In the above-mentioned benchmarks, no I was reported.

The correctness of localization for one sound source under six noise conditions was 100%, except for a small number of cases with up-TSP, down-TSP, and MN-SS: E5 errors occurred at 235° in up-TSP, at 40° in down-TSP, and at 335° in MN-SS.

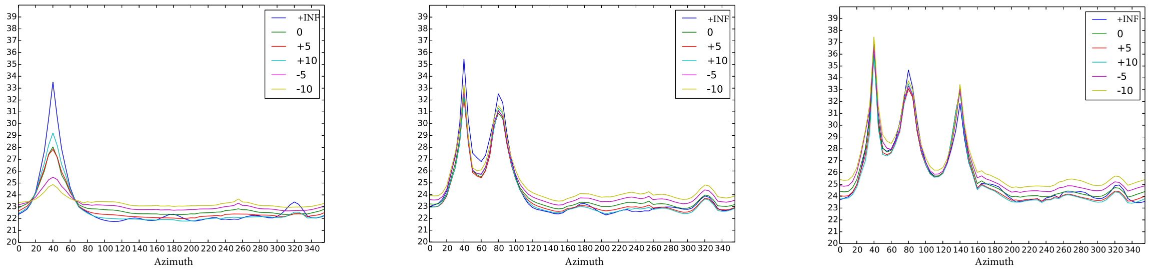
4.3. Two Sound Sources

The results of localization of two sound sources under six noise conditions are depicted in **Fig. 7**. Almost all sound sources were localized correctly except for a small number of cases with up-TSP, down-TSP, and MN-SS.

Most errors were E5 and E10 when their angles were less than or equal to 15°. Further, the missing leading part frequently occurred for narrower angles of two sound sources. Errors of E5 and E10 were approximately 2%–4% and 1%–2%, respectively. Missing leading part M was approximately 2% and the majority of the missing occurred within 10° of the angle between the two sound sources. Missing leading parts can be recovered by tuning the value of *THRESHOLD*; however, such tuning could cause other degradation.

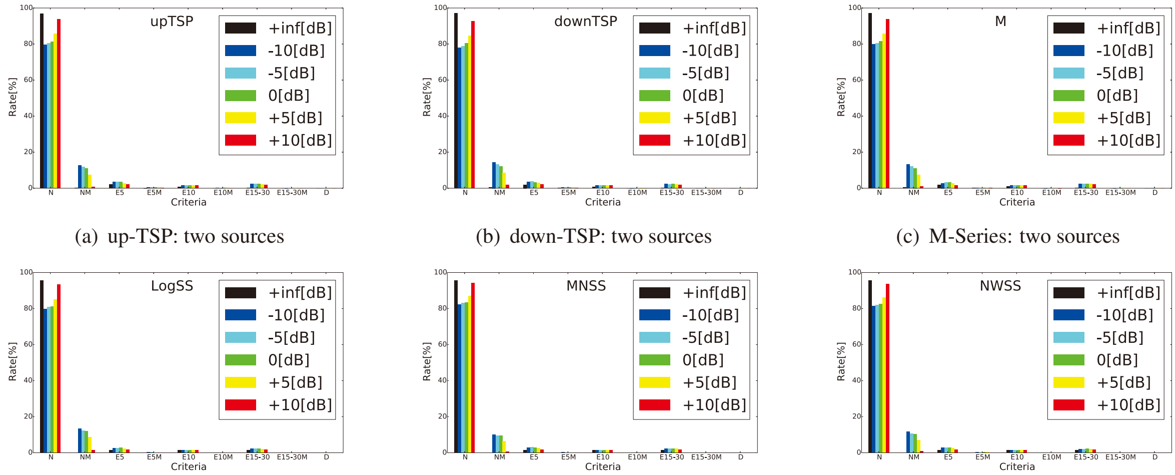
For comparing the influence of the six measurement signals, the rates of N and NM, the largest error, for different SNRs are summarized in **Fig. 8**. The influence on localization by the six measurement signals is rather minimal. To exaggerate the difference, MN-SS and NW-SS indicate slightly greater robustness for various combinations of sound sources in noisy situations, although they do not indicate higher performance in silent situations.

Figure 9 illustrates how the correctness, N and NM, of localization of two sound sources changes when their angle varies from 5° to 180° around the robot. It indicates



(a) one sound source (b) two sound sources (c) three sound sources

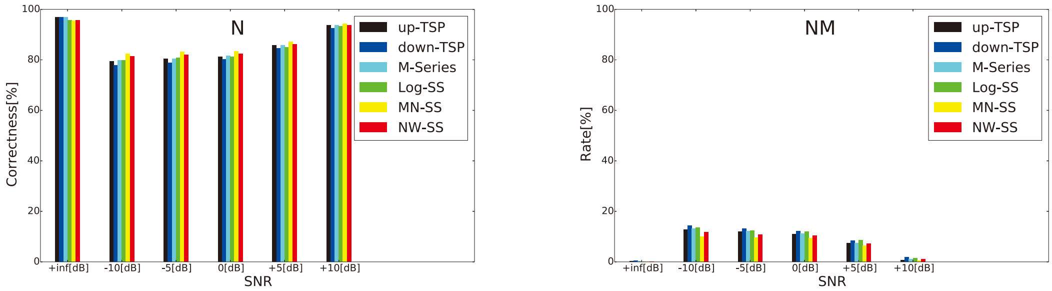
Fig. 6. MUSIC spectrogram of one to three sound sources with different noise levels.



(a) up-TSP: two sources (b) down-TSP: two sources (c) M-Series: two sources

(d) Log-SS: two sources (e) MN-SS: two sources (f) NW-SS: two sources

Fig. 7. SSL results of two sound sources for each measurement signal. First sound source is fixed at 0°; second moves from 5° to 355° by 5°. For each SNR level for each measurement signal, the data size is 10,224.



(a) Correctness (N) (b) NM rate

Fig. 8. Rates of N (correctness) and NM (correct but missing leading part) in localizing two sound sources.

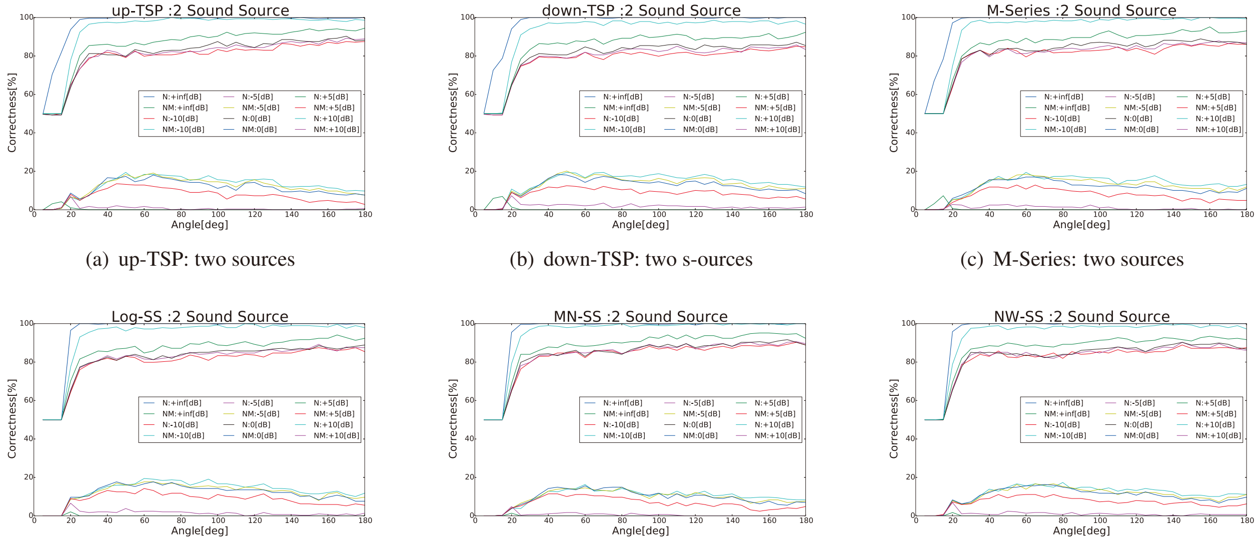
that HARK's MUSIC-based SSL can localize two sound sources effectively when their angle is more than 15°. For angles less than or equal to 15°, only one sound source is detected with impulse responses obtained by any of the six measuring signals.

4.4. Three Sound Sources

The correctness of localization of three sound sources under six noise conditions is depicted in **Fig. 10**. In general, no significant difference is observed. N is approximately 80% and errors of E15 and E10 are both at most 3%. NM is considerably smaller than in two sound sources. Some common symptoms were observed as follows:

- Errors were prone to occur at the azimuth of the errors in one sound source, for example, 235°.
- If two sound sources were localized as one sound source at their middle azimuth, for example, 45° for 40° and 50° sound sources, two E5 errors were recorded.
- Error of D was rather rare, that is, at most 0.3% for each case.
- The white noise level, even in ∞, did not influence the correctness in any measurement signals.

For comparing the influence of the six measurement signals, the rates of N and E5, the largest error, for different SNRs are summarized in **Fig. 11**. The influence on



(a) up-TSP: two sources

(b) down-TSP: two s-ources

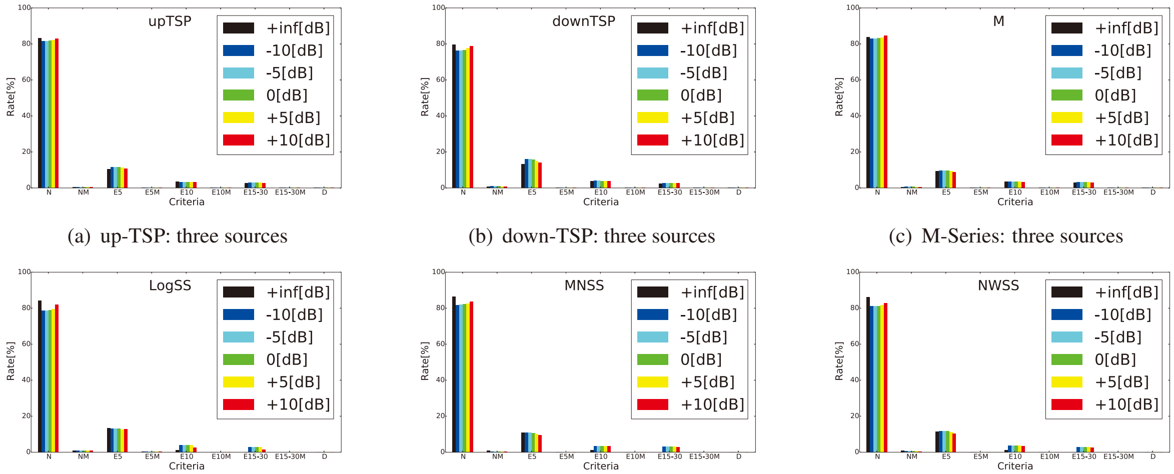
(c) M-Series: two sources

(d) Log-SS: two sources

(e) MN-SS: two sources

(f) NW-SS: two sources

Fig. 9. SSL correctness versus interval angle of two sound sources.



(a) up-TSP: three sources

(b) down-TSP: three sources

(c) M-Series: three sources

(d) Log-SS: three sources

(e) MN-SS: three sources

(f) NW-SS: three sources

Fig. 10. SSL results of three sound sources for each measurement signal. First sound sources is fixed at 0°, second moves to 5° to 350° by 5°, and third moves from the next 5° of the second to 355° by 5°. For each SNR level for each measurement signal, the data size is 1,073,520.

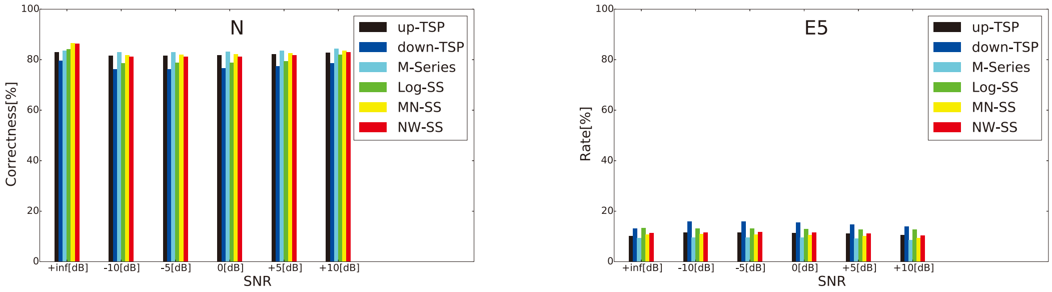
localization by the six measurement signals is minimal. To exaggerate the difference, M-Series, MN-SS and NW-SS indicate slightly greater robustness for various combinations of sound sources in noisy situations. One reason for this is that the benchmark sets of mixture of sounds that contain white noise favor MN-SS based on a minimum noise method and NW-SS based on a noise whitening method.

We focused on a particular pattern of positions of the three sound sources. The three sound sources moved around the robot by maintaining the same adjacent angle that changed from 5° to 120°. **Fig. 12** illustrates how the correctness of localization of three sound sources changed when their adjacent angular difference changed. It indicates that HARK's MUSIC-based SSL localizes three sound sources stably without being influenced by white noise. The correctness of localization becomes stable when their adjacent angular difference is more than or equal to 30°. Further, M or missing leading part is reduced drastically. This is important for SSS and ASR.

Based on the discussion on two sound sources, the performance of the MUSIC-based SSL of HARK is stable when their angular difference is more than 20°.

4.5. Further Analysis of SSL with Three Sound Sources

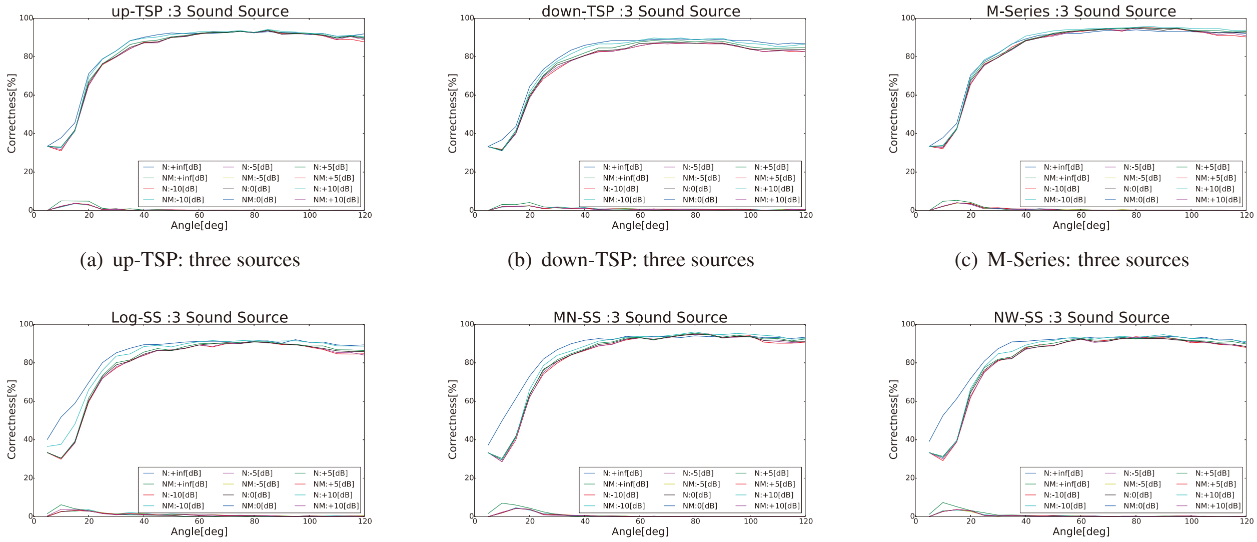
Suppose that the first and second sources are fixed at 0° and 20°, respectively, and the third moves from 40° to 340° by 5°. This pattern of positions reflects actual cases of three-person-party interactions. The azimuth-wise errors in the localization of the second and third sound sources are depicted in **Fig. 13** for each measurement signal. The observation is summarized in **Table 1**.



(a) Correctness (N)

(b) E5 rate

Fig. 11. Rates of N (correctness) and E5 (errors within ±5°) in localizing three sound sources.



(a) up-TSP: three sources

(b) down-TSP: three sources

(c) M-Series: three sources

(d) Log-SS: three sources

(e) MN-SS: three sources

(f) NW-SS: three sources

Fig. 12. SSL correctness versus the same interval angle of three sound sources. The interval angle varies from 5° to 120° by 5°.

This observation suggests the superiority of down-TSP and Log-SS, though the difference between them is minimal and some particular cases in **Figs. 9** and **12** demonstrate a tendency inconsistent with the suggested superiority.

As the spectrogram of the measured impulse responses of the six measurement signals presented in **Figs. 4(a)–(f)** in Subsection 3.2 indicates, the measured up-TSP has harmonic distortion from the beginning of the signal in the spectrogram, whereas neither down-TSP nor Log-SS indicates this. Because the transfer functions for MUSIC-based SSL of HARK uses only 32 points for impulse responses, harmonic distortion may degrade the performance of SSL. This superiority, although minimal, should be under scrutiny in the future.

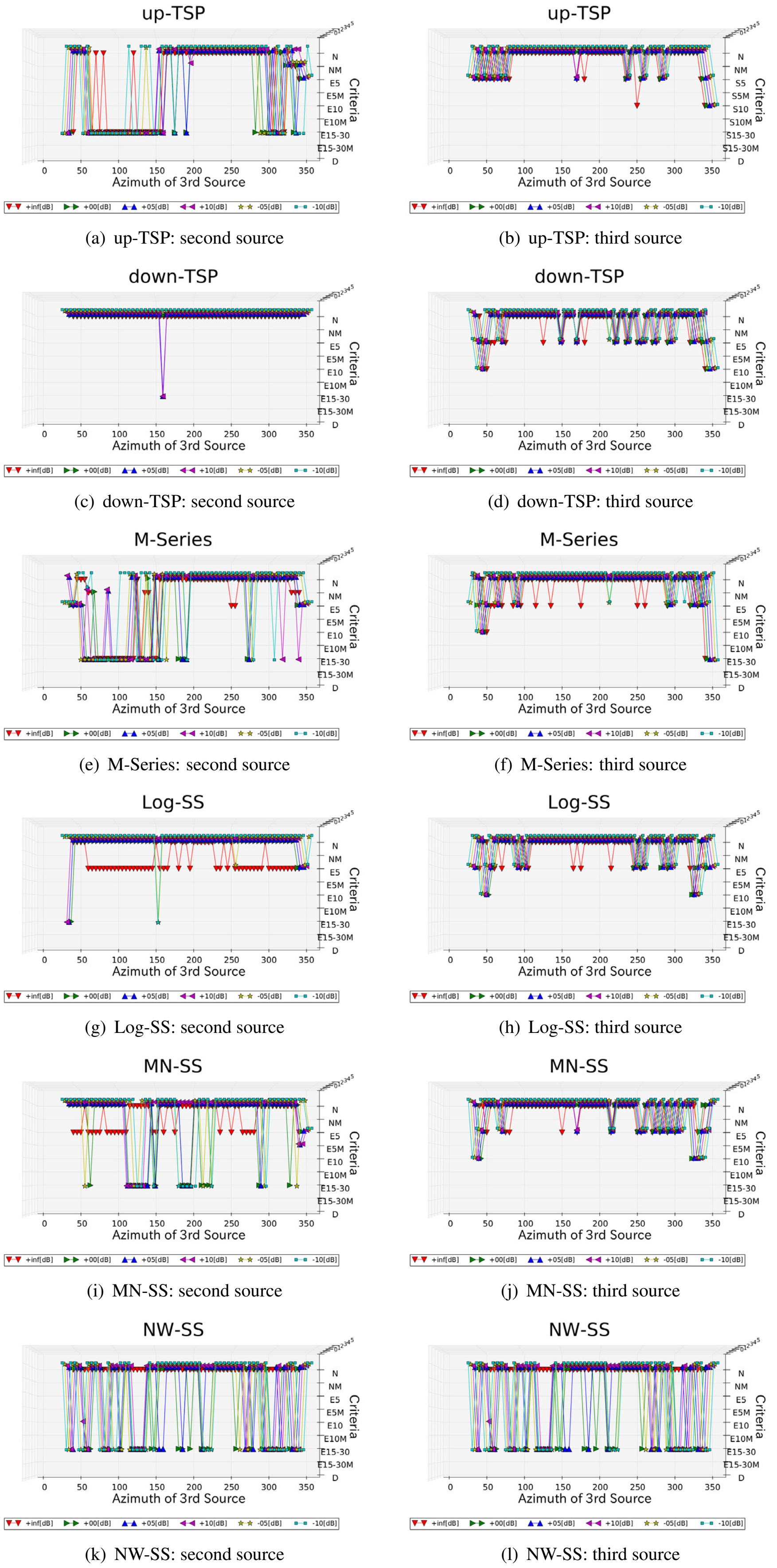
4.6. Limitations of Evaluation

Because the benchmark sounds use only a small number of utterances, the observations may depend on them. Although various kinds of techniques are applied to reduce the influence of noise, HARK tools do not adopt them. The next step of this paper is to evaluate the influence of the six measurement signals on SSS with ASR. Because SSS in HARK uses longer transfer functions, that is, 512 points, the influence should be estimated.

One of the most important remaining problems is to evaluate the influence of impulse measurement signals quantitatively using an “ideal” impulse response. An “ideal” impulse response can be obtained either by mathematical simulation of an acoustic field with a precise robot head model or by HARK’s impulse response calculation with the coordinates of a microphone array configuration. An evaluation with the latter method is in progress and its results will be reported in a separate paper.

5. Conclusion

In response to Kaneda’s appeal that the influence of an impulse measurement signal should be estimated from the viewpoint of the applications, this paper evaluated the influence of six measurement signals in terms of localization with the MUSIC-based sound source localization of HARK. Experimental results confirm no significant difference among the six measurement signals. Based on limited data of up to three simultaneous sound sources, down-TSP and Log-SS demonstrated slight superiority in the MUSIC-based localization. Because HARK provides pre-calculated transfer functions for commercially available microphones using up-TSP, the HARK community



(a) up-TSP: second source

(b) up-TSP: third source

(c) down-TSP: second source

(d) down-TSP: third source

(e) M-Series: second source

(f) M-Series: third source

(g) Log-SS: second source

(h) Log-SS: third source

(i) MN-SS: second source

(j) MN-SS: third source

(k) NW-SS: second source

(l) NW-SS: third source

Fig. 13. SSL results of three sound sources by six measurement signals in terms of criteria. First and second are fixed at 0° and 20°, respectively, and third moves from 40° to 340° by 5°.

Table 1. Summary of localization errors with 6 measurement signals.

Signal	Second Source	Third Source
common	No significant common observation.	· Errors were observed between 40° and 100° and between 250° and 300°. E10 were observed at around 40° and 340°. · White noise reduced errors.
up-TSP	E15-30 errors were observed between 60° and 150° and between 280° and 340°.	No E10 error was observed at around 40°.
down-TSP	Almost correct even at 20°.	No significant difference was observed among white noise levels.
M-Series	E15-30 errors were observed between 60° and 150°.	White noise reduced errors between 100° and 300°. E15-30 errors were observed at 340°.
Log-SS	E5 errors were observed without white noise, whereas white noise reduced most errors.	No significant observation.
MN-SS	E5 errors were observed without white noise, whereas white noise caused E15-30 errors.	White noise influence was small.
NW-SS	Without white noise, no error was observed, whereas white noise caused many E15-30 errors.	No E10 errors were observed around 40°.

should reexamine the use of other impulse response measurement signals instead of up-TSP for improving the performance of SSL. Their contribution to the improvement of SSS is an interesting open problem.

Acknowledgements

The authors appreciate Professor Yutaka Kaneda for his valued supports and Dr. Kazuhiro Nakadai for his valuable advice and discussion. This research is supported primarily by JSPS Kakenhi (S) No.2420006, and partially supported by ImPACT Tough Robotic Challenge and Waseda University Leading Graduate Program for Embodiment Informatics.

References:

[1] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active Audition for Humanoid," Proc. of the Seventeenth National Conf. on Artificial Intelligence (AAAI-2000), pp. 832-839, 2000.

[2] K. Nakadai, H. G. Okuno, and H. Kitano, "Real-Time Auditory and Visual Multiple-Speaker Tracking For Human-Robot Interaction," J. of Robotics and Mechatronics, Vol.14, No.5, pp. 479-489, 2002.

[3] I. Nishimuta, K. Yoshii, K. Itoyama, and H. G. Okuno, "Toward a Quizmaster Robot for Speech-based Multiparty Interaction," Advanced Robotics, Vol.29, Issue 18, pp. 1205-1219, Sep. 2015.

[4] H. G. Okuno and K. Nakadai, "Robot Audition: Its Rise and Perspectives," Proc. of 2015 Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2015), pp. 5610-5614, 2015.

[5] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and Implementation of Robot Audition System "HARK" – Open Source Software for Listening to Three Simulteaneous Speakers," Advanced Robotics, Vol.24, Issue 5-6, pp. 739-761, Jan. 2010.

[6] K. Nakamura, K. Nakadai, F. Asano, and G. Ince, "Intelligent Sound Source Localization and Its Application to Multimodal Human Tracking," Proc. of 2011 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 2011), pp. 143-148, 2011.

[7] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," IEEE Trans. on Antennas and Propagation, Vol.34, No.3, pp. 276-280, 1986.

[8] F. Asano, M. Goto, and H. Aso, "Real-time Sound Source Localization and Separation System and Its Application to Automatic Speech Recognition," Proc. of EUROSPEECH-2001, pp. 1013-1016, 2001.

[9] S. Müller, "Measuring Transfer-Functions and Impulse Responses," Chapter 5, Springer Handbook of Acoustics, p. 1000, Springer, 2009.

[10] S. Weinzierl, A. Giese, and A. Lindau, "Generalized multiple sweep measurement," Proc. of 126th AES Convention, p. 7767, 2009.

[11] P. Majdak, P. Balazs, and B. Laback, "Multiple Exponential Sweep Method for Fast Measurement of Head-Related Transfer Functions," J. of Audio Engineering Society, Vol.55, Issue 7/8, pp. 623-637, 2007.

[12] Y. Kaneda, "Measurement signals for an acoustical impulse response," invited talk, IEICE Technical Report, EA2015-68, SIP2015-117, SP2015-96, Mar. 2016 (in Japanese).

[13] K. Nakamura, K. Nakadai, and H. G. Okuno, "A real-tome super-resolution robot audition system that improves the robustness of simultaneous speech recognition," Advanced Robotics, Vol.27, Issue 12, pp. 933-945, 2013.

[14] G.-B. Stan, J.-J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," J. of Acoustic Society of America, Vol.50, No.4, pp. 249-262, Apr. 2002.

[15] M. R. Schroeder, "Integrated-impulse method for measuring sound decay without using impulses," J. of Acoustic Society of America, Vol.66, No.2, pp. 497-500, 1933.

[16] M. R. Schroeder, "Number Theory in Science and Communication," Spriger-Verlag, 1984.

[17] N. Aoshima, "Computer-generated pulse signal applied for sound measurement," J. of Acoustic Society of America, Vol.69, No.5, pp. 1483-1488, 1981.

[18] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," J. of Acoustic Society of America, Vol.97, No.2, pp. 1119-1123, 1995.

[19] S. Müller and P. Massarani, "Transfer Function Measurement with Sweeps," J. of Audio Engineering Society, Vol.49, No.6, pp. 443-471, 2011.

[20] A. Farina, "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Techniques," Proc. of 108th Convention of Audio Engineering Society, p. 5093, Paris, February 2000.

[21] N. Moriya and Y. Kaneda, "Optimum signal for impulse response measurement that minimizes error caused by ambient noise," J. of Acoustic Society of Japan, Vol.64, No.12, pp. 695-701, 2008 (in Japanese).

[22] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," IEEE Signal Processing Magazine, Vol.13, No.4, pp. 67-94, 1996.

[23] K. Nakamura, K. Nakadai, F. Asano, and H. Tsujino, "Intelligent Sound Source Localization for Dynamic Environments," Proc. of 2009 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 2009), pp. 664-669, 2009.

[24] W. Akahori, T. Masuda, H. G. Okuno, and S. Morishima, "The Evaluation of influence of Measurement Methods of the Transfer Function on Sound Source Localization and Separation," Proc. of the 77th Annual Meeting of Information Processing Society of Japan, 5P-03, pp. 119-120, Mar. 2015.

[25] K. Ito, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition," J. of Acoustic Society of Japan (E), Vol.20, No.3, pp. 199-206, 1999.

[26] T. Takahshi, K. Nakadai, C. T. Ishi, and H. G. Okuno, "Investigation of Sound Source Localization and Separation under a Real Environment," Proc. of 29th Annual Meeting of Robotics Society of Japan, AC1EF3-3, 2011 (in Japanese).

Supporting Online Materials:

[a] HARK group, "6.2.1 LocalizeMUSIC," HARK Document, Version 2.2.0 (Revision: 7981).
http://www.hark.jp/document/hark-document-en/
subsec-LocalizeMUSIC.html [Accessed July 20, 2016]

Name:
Takuya Suzuki

Affiliation:
Waseda University

Address:
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

Brief Biographical History:
2015- Part-time Research Assistant, Graduate Program for Embodiment Informatics, Waseda University
2016- Junior of Department of Computer Science and Engineering, Waseda University

Main Works:
- robot audition and robotic musicianship

Membership in Academic Societies:
- Information Processing Society of Japan (IPSJ)

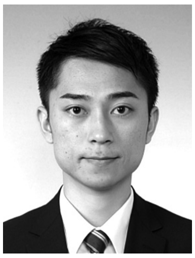
Name:
Hiroaki Otsuka

Affiliation:
Waseda University

Address:
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

Brief Biographical History:
2015- Part-time Research Assistant, Graduate Program for Embodiment Informatics, Waseda University
2016- Junior of Department of Electrical Engineering and Bioscience, Waseda University

Main Works:
- robot audition



Name:
Wataru Akahori

Affiliation:
Waseda University

Address:
55N406, 3-4-1 Okubo, Shinjuku-ku, Tokyo 161-8555, Japan

Brief Biographical History:
2015 Received Bachelor of Engineering (B.E) from Waseda University
2015- Master of Engineering (M.E.) Student, Waseda University

Main Works:
- user experience design

Membership in Academic Societies:
- Information Processing Society of Japan (IPSJ)
- Association for Computing Machinery (ACM)

Name:
Yoshiaki Bando

Affiliation:
Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University
Research Fellowship for Young Scientists (DC1), Japan Society for the Promotion of Science

Address:
Room 417, Research Bldg. No.7, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan

Brief Biographical History:
2014 Received M.Inf. degree from Graduate School of Informatics, Kyoto University
2015- Ph.D. Candidate, Graduate School of Informatics, Kyoto University

Main Works:
- "Posture estimation of hose-shaped robot by using active microphone array," Advanced Robotics, Vol.29, No.1, pp. 35-49, 2015 (Advanced Robotics Best Paper Award).

Membership in Academic Societies:
- The Institute of Electrical and Electronic Engineers (IEEE) Robot Automation Society (RAS)
- The Robotics Society of Japan (RSJ)
- Information Processing Society of Japan (IPSJ)

Name:
Hiroshi G. Okuno

Affiliation:
Professor, Graduate School of Science and Engineering, Waseda University
Professor Emeritus, Kyoto University

Address:
Lambdax Bldg 3F, 2-4-12 Okubo, Shinjuku, Tokyo 169-0072, Japan

Brief Biographical History:
1996 Received Ph.D. of Engineering from Graduate School of Engineering, The University of Tokyo
2001-2014 Professor, Graduate School of Informatics, Kyoto University
2014- Professor, Graduate School of Science and Engineering, Waseda University

Main Works:
- "Design and Implementation of Robot Audition System "HARK"," Advanced Robotics, Vol.24, No.5-6, pp. 739-761, 2010.
- "Computational Auditory Scene Analysis," Lawrence Erlbaum Associates, Mahmoh, NJ, 1998.

Membership in Academic Societies:
- The Institute of Electrical and Electronic Engineers (IEEE), Fellow
- The Japanese Society for Artificial Intelligence (JSAI), Fellow
- Information Processing Society Japan (IPSJ), Fellow
- The Robotics Society of Japan (RSJ), Fellow