A sound-selective hearing support system using environment sensor network

Carlos Toshinori Ishi1,*, Chaoran Liu1,†, Jani Even1,‡ and Norihiro Hagita2,§

1ATR HIL, 2–2–2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619–0288 Japan
2ATR IRC, 2–2–2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619–0288 Japan

(Received 5 July 2017, Accepted for publication 29 December 2017)

Abstract: We have developed a sound-selective hearing support system by making use of an environment sensor network, so that individual target and anti-target sound sources in the environment can be selected, and spatial information of the target sound sources can be reconstructed. The performance of the selective sound separation module was evaluated under different noise conditions. Results showed that signal-to-noise ratios of around 15 dB could be achieved by the proposed system for a 65 dB babble noise plus directional music noise condition. Subjective intelligibility tests were conducted in the same noise condition. For words with high familiarity, intelligibility scores increased from 67% to 90% for normal hearing subjects and from 50% to 70% for elderly subjects, when the proposed system was applied.

Keywords: Hearing support, Intelligibility, Sound separation, Sensor network

PACS number: 43.66.Qp, 43.60.Dh, 43.66.Ts [doi:10.1250/ast.39.287]

1. INTRODUCTION

According to WHO (World Health Organization), over 5% of the world’s population (360 million people) has disabling hearing loss [1]. In Japan, it is reported that 11.3% of the population has hearing loss (self-declared), from which only 1.5% are hearing aid device users [2]. From those, 40% declared not being satisfied with the current devices. Further, most of the hearing aid device users stop using it, mainly because unwanted disturbing sounds are also amplified along with the target sounds, e.g. the dialogue partner speech, so that they cannot have normal dialogue communication in their daily life. Another problem is that spatial information of sounds is lost.

In order to solve the problems of current hearing aid devices, we aim on developing a hearing support system which is able to select necessary sounds (such as dialogue partner’s voice, bells, TV sounds) and unnecessary or disturbing sounds (such as door, air conditioner, voices of people other than dialogue partners), and can recover spatial information of the sounds, by making use of sound environment intelligence technologies.

There are several works applying binaural signal processing (producing output for both ears) to improve the performance of hearing aids. The goal is not only to extract the desired speech and suppress the background noise but also to preserve the binaural cues in order to preserve the auditory impression [3]. For example, many works apply beamforming and multi-channel Wiener filtering (e.g. [4,5]). Others apply blind source separation and post-filtering techniques to improve the separation of the target sound source [6,7]. The use of sensor network to incorporate additional microphones (usually placed in the room) has also been investigated [8,9]. However, most of works apply binaural processing using the microphones of the hearing aid or a microphone array worn by the user [10].

In this work, we propose a hearing support system, which makes use of environment sensors, and allows the selection of multiple sound sources in the environment. This contrasts with most works that emphasize one sound source, in the look direction of the user.

In our previous studies, we have proposed a method for estimating the sound source positions in 3D space by combining the direction estimations of multiple microphone arrays [11], and combined the information of sound source positions and human positions, in order to discriminate between speech and non-speech sound events, so that information about speech activity for each human in the environment is obtained [12].
In the present work, we make use of such sound environment intelligence technologies, and developed a prototype of a hearing support system, where individual target and anti-target sound sources in the environment can be selected, and spatial information of the target sound sources is reconstructed.

The performance of the sound separation module was evaluated each step, based on SNR (signal to noise ratio) improvements. We then evaluated the performance of the whole system throughout subjective listening test in noise conditions, done by people with normal hearing (20 s–50 s) and elderly people (in their 70 s). The present manuscript is an extension of our previous report in [13].

2. THE PROPOSED SYSTEM

Figure 1 shows a block diagram of the proposed hearing support system. The proposed system is composed by two parts. One is the environment sensor network side, where sound direction and human position information are integrated, and individual sound sources are separated. The other is the user side, where target and anti-target sound sources are selected, and the user’s head orientation is tracked to reconstruct the spatial information of the target sounds.

The basic structure of the present system is similar to our previously proposed teleoperation robot system, where acoustic sense of presence is transmitted from the robot’s environment to the tele-operator side [14]. One difference to the previously reported system is that the tele-operator is in a remote place in the teleoperation system, while the user shares the same place of the sensor network in the hearing support system, so that the sensor network has to be aware of the user’s position and orientation. Another difference is that, in the present work, the sound separation module was considerably improved by including multi-channel background noise suppression and inter-channel suppression modules (Sect. 2.2).

In the environment sensor network side, directions of multiple sound sources are firstly estimated in the 3D-space (i.e., both azimuth and elevation angles) for each microphone array. The estimated sound directions are integrated in the 3D room geometry space with the human positions provided by the human tracker module, to provide who is talking when and where in the environment [12]. The sound source locations where human positions are not detected can be attributed to non-human sound sources. Then, individual sound sources are separated and sent to the user side, along with their positions. If multiple arrays are available, the output of the array closest to the sound source is selected.

In the user side, the 3D positions of the sound sources and the user’s head orientation, measured from the gyro/compass sensor attached to the user’s headphone, are used to obtain the relative angles between the user and the sound sources. These angles are then used to select binaural HRTF (Head-Related Transfer Functions [15]) from a database. Each separated signal is convolved with the corresponding left and right ear HRTFs, summed out to reconstruct the auditory scene (spatial sensation of sounds), and played in the user’s headphones. The volume of each sound source is optionally controllable through a user interface.

For the human tracking, we used a particle filter-based 2D human tracking using 2D LRF (Laser Range Finders) [16]. The human position data is obtained in approximately 33 ms resolution.

For the 3D-space sound direction estimation, we use our previously proposed method based on the MUSIC (MUltiple SIgnal Classification) algorithm [17]. A hierarchical approach is also adopted, in order to obtain direction estimations with 1 degree resolution in the 3D-space, each 100 ms, in real-time by a 2.6 GHz CPU core [12].

In the following subsections, details about the spatial information integration, sound separation and auditory scene reconstruction methods are explained.

2.1. Spatial Information Integration

Sound source activities are detected by integrating spatial information of sound directions (estimated by the
As a first step, background noise suppression is conducted for each channel separately, to reduce stationary noises (such as air conditioner and PC fan noises). For the conducted for each channel separately, to reduce stationary processing is conducted for all selected sound sources.

### 2.2. Sound Source Separation Processing

For the sound source separation processing, parallel processing is conducted for all selected sound sources. Figure 2 shows the processing flow.

As a first step, background noise suppression is conducted for each channel separately, to reduce stationary noises (such as air conditioner and PC fan noises). For the background noise suppression (BNS), Wiener filter $H_{\text{BNS}}$ is computed for each channel $i$ and frequency bin $f$, as in the following expression.

$$H_{\text{BNS}}(f) = \frac{1}{1 + \frac{N_i(f)}{X_i(f)}}$$

The noise estimation ($N_i(f)$) is conducted by averaging the spectrum in the intervals where target sounds are absent. In practice, the noise estimation is conducted in the first few seconds, right after the system is launched.

Although the noise suppression is commonly conducted as a post-processing, after the beamforming, it is conducted prior to the beamformer in the present work, in order to reduce the effects of musical noise in the separated signal. ("Musical noise" is a random tonal noise which arises as an artefact of non-linear noise suppression processing. Musical noise components differ from channel to channel, so that after beamforming, these components will be smoothed over the whole channels.)

Next, beamforming is conducted for each selected sound source, based on the directions (azimuth $\theta$ and elevation $\phi$) and range information $(r)$, obtained from the spatial information integration module. In the present work, we use the DS (Delay-Sum) beamformer due to its robustness, low computational costs, and low latencies. Frame length of 32 ms and frame shift of 10 ms is used for beamforming.

As a constraint of the size of the microphone array, the DS beamformer has low resolution for low frequencies. As a result, low frequency components of the noise and interference sources will remain in the separated signal.

Considering the target directional sound source $S$ and the diffuse noise sources $N$, the DS beamformer output can be written in the following form.

$$Y_{\text{DS}}(f) = w_{\text{DS}}(f) \cdot S(f) + \int_0^{2\pi} \left( w_\theta(f) \cdot N(f) \right) d\theta$$

(2)

$Y_{\text{DS}}(f)$ is the beamformer output for frequency $f$, $S_{\text{dir}}$ is the direction of the signal, $w_{\text{DS}}$ are the beamformer weights for the direction $S_{\text{dir}}$, and $w_\theta$ are the steering vectors for the direction $\theta$. The second item in the right side of Eq. (2) corresponds to the noise components remained in the separated signal. In order to reduce the effects of the low frequency noise components, each frequency component was multiplied by the following weight function.

$$w_{\text{norm}}(f) = \frac{1}{\int_0^{2\pi} w_\theta(f) d\theta}$$

(3)

$$Y_j(f) = w_{\text{norm}}(f) \cdot Y_{\text{DS}}(f)$$

(4)

$Y_j$ is the beamformer output of the $j$-th sound source, after multiplying the weighting functions.
The DS beamformer emphasizes a signal in a specified direction, but the suppression gain for signals in other directions is not high. In order to improve the separation of directional sound sources, inter-channel suppression (ICS) is conducted. The interference sources are suppressed by Wiener filtering \( H_{ICS} \), according to the following expressions.

\[
H_{ICS}(f) = \frac{1}{1 + \frac{I_i(f)}{Y_i(f)}}
\]  

(5)

\[
I_i(f) = \max_{j \neq i} |Y_j(f)|
\]  

(6)

\( I_i(f) \) is the maximum spectral component among the \( J \) directional sound sources excluding the \( i \)-th target source.

One problem related to the directional interference source suppression is that, if the angle between the directions of the target source and interference sources are small, the beamforming performance will be low, so that inter-channel suppression would produce a highly distorted output signal. Thus, in order to avoid such distortion, we included a constraint for the suppression processing, according to the following expression.

\[
I_i(f) = \frac{|dir_1 - dir_2|}{5} I_i(f), \text{ if } |dir_1 - dir_2| < 5
\]  

(7)

In this way, suppression levels will be reduced when the angle between target and interference sources are smaller than 5 degrees. We verified that for angles larger than 5 degrees, the performance of the DS beamformer was high enough by our 16-channel microphone array.

Finally, gain normalization is conducted to compensate the sound pressure attenuation of the observed signal, according to the distance between the array and the sound source, according to the following expression.

\[
g_i = \begin{cases} 
0.5, & r_i < 0.5 \\
2, & r_i > 2 \\
2, & 0.5 < r_i < 2 
\end{cases}
\]  

(8)

\[
Sep_i = g_i \cdot Y_i
\]  

(9)

\( g_i \) is the normalization gain, \( r_i \) is the range between the array and the \( i \)-th source in meters, and \( Sep_i \) is the separated signal for the \( i \)-th source. The constraints of 0.5 and 2 m for the range were imposed since preliminary evaluation indicated that the background noise levels increased too much for large ranges.

2.3. Auditory Scene Reconstruction

The auditory scene (spatial sensation of sound sources) is reconstructed for each separated sound received from the sensor network side, considering the positioning of the sound sources relative to the user. The auditory scene is reconstructed using HRTF (Head-related transfer functions) for each sound source.

In the present work, we use the freely available HRTF database of the KEMAR (Knowles Electronics Manikin for Acoustic Research) dummy head microphones [18]. The database consists of the left and right ear impulse response measurements from a loudspeaker mounted 1.4 meters from the KEMAR, at a sampling rate of 44.1 kHz. A total of 710 different positions were sampled at elevations from −40 degrees to +90 degrees, each 5 degrees.

In order to dynamically reconstruct the auditory scene in real-time, the user’s head orientation has to be tracked with low latency. For that purpose, a gyro/compass sensor was attached on the top of the headphone to track the head rotation angles. The angle information is sent through serial communication to the user’s device. The relative direction of the sound source is then obtained by subtracting the head angle from the angle provided by the sensor network in global coordinates. The left and right channel impulse responses with the closest angles are selected from the database, and convolved with the separated source signals. The resulting stereo signals are finally played to the user’s headphones.

3. ANALYSIS DATA DESCRIPTION AND EVALUATION OF THE PROPOSED METHOD

3.1. Experimental Setup

The evaluation experiments were conducted in a sound-proof room, where noise sources were generated by multiple loudspeakers, as shown in Fig. 3.

Two types of noise source were generated. One is a babble noise (mixture of voices by multiple non-target people) with sound pressure adjusted to 60–65 dB. The loudspeaker for the babble noise was placed under the table, in order to simulate a non-directional (diffuse) noise, relative to both subject and microphone array. The other noise source is a background ambient music (English pop music), with sound pressure adjusted to around 60 dB. The loudspeaker for the background music was placed at around 1 m from the array, in order to simulate a directional noise source. The same song was played.
repeatedly. The loudspeaker for the target speech source was placed around 1.5 m in front of the subject, with average sound pressure adjusted to 60–65 dB. In this way, SNR (signal-to-noise ratio) around 0–5 dB were simulated.

For the target speech, we used two sets of word lists with different degrees of familiarity [19]. The first set (‘Fam70’) is composed by 50 words with relatively high familiarity degrees (55–70%), while the second set (‘Fam25’) is composed by 50 words with low familiarity degrees (10–25%). All words are composed by four morae (‘mora’ is a metrical unit equivalent to a short syllable). The words were read by an adult female speaker.

For all loudspeakers, we used Yamaha MS101III. For the headphone to play the output signal of the proposed system, we used the Bose QuietComfort 25, which has active noise canceller (ANC) function. The ANC function was activated during the experiments in order to reduce the levels of the sounds coming directly from the sound sources (i.e., from outside the headphones). This is important to ensure that the levels of the delayed (processed) sounds which comes through the headphone speakers are sufficiently louder than the (unprocessed non-delayed) sounds which leaks from outside the headphones. Preliminary subjective tests indicated that listenability improves by using the ANC function.

Figure 4 shows the geometry of the 16-channel microphone array on the table. The microphones are distributed over a half-sphere of 30 cm diameter, in order to provide direction estimation in the 3D space, i.e. estimation of both azimuth and elevation angles [12]. For the multi-channel audio capture device, we used the 16-channel A/D converter TD-BD-16ADUSB from the Tokyo Electron Devices Ltd. For the microphones, we used the Sony omnidirectional condenser microphones ECM-C10. Audio was sampled at 16 kHz/16 bits resolution for all 16 channels.

For the parameter setting of the sound direction estimation based on the MUSIC method, the number of sources for the MUSIC spectrum computation was fixed to 3, the threshold for selecting peaks in the MUSIC spectrum was set to 2.5 dB, and the maximum number of simultaneous sources at the same block was set to 6. The frequency range of operation for estimating the MUSIC spectrum was set to 1,000–5,000 Hz, for avoiding spatial aliasing in high frequencies, and low spatial resolution in low frequencies. For the search space of sound direction estimation, a range of 0–360 degrees is set for azimuth angle, while a range of 10–80 degrees is set for elevation angle.

For the Wiener filter parameters in both background noise suppression (BNS) and inter-channel suppression (ICS) modules, an attenuation factor $\alpha = 1$ and a noise floor factor $\beta = 0.001$ were applied.

For the human position estimation, we used two 2D laser range finder (LRF) sensors, placed in the room corners. For the LRFs, we used UTM-30L from Hokuyo Automatic Co. Ltd. The height of the LRF sensors was adjusted to about 1.1 m, in order to be able to detect the positions of the subject and the loudspeakers. In the user interface display, the positions of the sound sources are shown, so that the user can select the speech source as a target, and the music source as an anti-target. The babble noise source cannot be explicitly selected in the present experiment, since it is used to simulate a diffuse noise, and it cannot be seen by both LRF and microphone array sensors.

The detected source heights shown in Fig. 3 were 1.11 ± 0.05 m for the target loudspeaker, 1.08 ± 0.02 m for the music loudspeaker, 1.12 ± 0.04 for one of the participants, and 1.09 ± 0.05 for the experimenter. Given that the true positions of all sound sources are around 1.1 m, estimation errors are within a few centimeters, indicating the feasibility of the spatial information integration module.

3.2. Evaluation of the Proposed Sound Source Separation Module

In order to evaluate the effects of the different components of the sound separation module, the following processing conditions were evaluated: DS beamformer only (‘DS”), background noise suppression followed by DS beamformer (“BNS+DS”), DS beamformer followed by inter-channel suppression (“DS+ICS”), and all processing (“BNS+DS+ICS”).

In order to evaluate the effects of each noise type, the following noise conditions were evaluated: clean environment (“clean”), diffuse babble noise only (“babble”), directional music noise only (“music”), babble noise + music noise (“babble+music”).

SNRs are computed at the frame level and averaged per utterance. Figure 5 shows the SNRs (mean and standard deviations of the utterance SNRs) for each separation processing condition and each noise condition. “RAW”
represents the results for unprocessed raw signal in one of the microphones in the array.

The “clean” environment condition actually contained a small fan noise of the note PC used in the experiments. In that case, the SNR was around 20 dB, as shown in the first bar of Fig. 5.

From the music noise, it can be observed that SNR improves by both background noise suppression (BNS) and inter-channel suppression (ICS) processing. However, for the babble noise, it can be observed that the inclusion of background noise suppression does not improve the SNR. This is because the babble noise was generated to be diffuse, and the noise estimation was conducted before the appearance of babble noise, in order to avoid distortion of the separated signals. For the “babble+music” noise condition, the SNR of the separated signals was around 15 dB, after inter-channel suppression (ICS). Considering that the acceptable noise level (ANL) for both normal-hearing and hearing-impaired people is around 10 dB [20], one can expect high intelligibility by using the proposed system.

### 3.3. Evaluation of the Proposed System through Listening Tests

The effectiveness of the proposed hearing support system was evaluated through intelligibility tests conducted in the “babble+music” noise condition (which is the most severe condition). We first evaluated the performance of the system by normal hearing people (self-reported), and then conducted evaluation experiments by elderly people (with high probability of having hearing loss).

For the listening tests, we used two sets of Japanese word lists with different degrees of familiarity (“Fam70” and “Fam25”), as described in Sect. 3.1.

Ten people with normal hearing (male and female, aging from 20s to 50s), and ten elderly people (male and female aging at their 70s) participated in the experiments. Although none of the elderly participants wear hearing aid devices, some of them reported that they have to set a higher volume while watching TV.

The listening tests were conducted firstly without the system (“without system”), and later with the proposed system by wearing the headphones (“proposed system”). Although the same word sets (100 words in total) were used in the two experiments, learning effects are considered to be small since the number of words is big, and the subjects listened to the same word after about 10 minutes.

The timing for playing the target words was controlled by the experimenter, in a way to play the next word after the participant finishes writing the previous word.

For the proposed system, four sound sources are selected. The target speech loudspeaker and the experimenter are selected as targets, while the music loudspeaker and the user him/herself are selected as anti-targets. Although there is no dialogue interaction during the listening tests, the user was also selected as an anti-target, since during conversations the user’s voice would be played-back with some delay, which would cause difficulty in speaking.

Figure 6 shows the intelligibility results for the different conditions, for normal hearing subjects (left panel) and elderly subjects (right panel). Intelligibility scores were computed at word and mora levels.

From the results in Fig. 6, it can be observed that the intelligibility scores increased by the use of the proposed system compared to not using it, in all conditions, for both normal hearing and elderly subject groups (all differences were statistically significant with \( p < 0.01 \) by \( t \)-tests).

For the words with high familiarity, word intelligibility scores increased from 67% to more than 90% for normal hearing subjects, by using the proposed system. For elderly subjects, the word intelligibility increased from 50% to 70% on average for words with high familiarity.

For the words with low familiarity, the word intelligibility improved, but the average correct rate is still low (around 50–60%). However, the mora intelligibility scores are higher (around 80–90%). The reason for these results is that most of listening errors were on average one mora per word.
3.4. Evaluation of the Hearing Support System through Subjective Impression

For the elderly subjects, after the listening tests, free conversations were conducted between the subject and the experimenter in the babble+music noise condition, with the following three sessions: 1) without the system, 2) with the system (monaural: same signal to the both ears, prior to the HRTF filters), and 3) with the system (binaural: left and right signals after HRTF filters). The conversations lasted approximately three minutes per session.

The subjects are asked to answer their subjective impressions according to the following questionnaire.

Which noise source was more annoying? (7-point scale from $-3$ for babble noise to 3 for music noise.)

Which condition was better for listenability? (7-point scale, from $-3$ for “without the system” to 3 for “with the proposed system.”)

Regarding the “proposed system,” subjective impression was scored in 4-point scales, for listenability (1: very difficult; 2: slightly difficult; 3: easy; 4: very easy), comfortability (1: uncomfortable; 2: acceptable; 3: comfortable, 4: very comfortable), annoyance about the musical noise and the delay (1: not annoying at all; 2: slightly annoying; 3: annoying; 4: very annoying), and preference between monaural and binaural modes.

Figure 7 shows the subjective impression results. Firstly, regarding the noise type, some of the subjects judged the babble noise to be more annoying, but most of subjects answered that both noise types were annoying. Regarding the comparative listenability, almost all subjects answered that they prefer to use the system.

Regarding the subjective impressions about the system, most of subjects judged the system to be easy to listen, and acceptable comfortability. For the musical noise and delay, more than half of the subjects judged both as not to slightly annoying.

Finally, for the comparison between monaural and binaural modes, nine of the ten subjects preferred binaural mode. The subject that preferred the monaural mode answered that the high frequency sound was annoying. This is probably because the HRTFs emphasize the high frequency components, comparing to the monaural signals.

4. DISCUSSION

From the results in Sect. 3.3, it can be observed that for words with high familiarity, the intelligibility by elderly people using the system (70%) is higher than by normal hearing people not using the system (67%). For the words with low familiarity (which are close to meaningless words), the intelligibility scores increased but are still low for both elderly and normal hearing people. These results suggest that the system is effective for providing hearing support with word intelligibilities equivalent or slightly superior to normal hearing, for daily conversation situations where words with high familiarity are predominant.

Regarding the practicability of the proposed system, the device worn by users could be switched from the normal use (using the microphones embedded in the user device) to the sensor network use (proposed system), when the user is within a region where the sensor network is available. For the environment sensor network, although a single microphone array was used in the present evaluation, the number of arrays can be increased to enlarge the target source coverage area. In that case, the array sensors should be arranged in order to be as “visible” as possible by the target sound sources (i.e., with few occlusions).

The system delay is another issue for practical use, since most of commercial hearing aids have delays of the order of 10 ms. On the other hand, recent works have reported that hearing-impaired adults have a higher tolerance for larger delays (40–80 ms) when speech is presented in noisy environments [21]. Although the delay of the system was set to 200 ms in the experiments to ensure that the audio signal is not cut when eventual delays occur in the data transmission, the actual processing latency of the current implementation is around 60 ms. For practical use, some efforts are still necessary to reduce this latency, for example by implementing filters in time domain, instead of the current implementation in frequency domain.

The sound sources in the present work were simulated by loudspeakers which were identified by the human tracker module. However, in practice, non-human sound sources can be dealt with in the same way, if their locations are provided. For example, most of noise sources (such as air conditioner, TV, loudspeakers, doors, PCs, bells, etc.) are in specific locations in an environment. In that case, the user can preset the locations of the target and anti-target sounds, so that the necessary sounds in daily-life are emphasized and the unnecessary sounds are suppressed. Mobile robots can also be used to create 3D sound environment maps [22], in order to provide prior knowledge about sound source locations.
5. CONCLUSIONS

In the present work, we developed and evaluated a hearing support system by applying sound environment intelligence technologies, where individual sound sources are separated in the sensor network side, and a selection of target and anti-target sounds with auditory scene reconstruction is conducted in the user side.

The performance of the sound separation module was evaluated under different noise conditions. The results showed that signal-to-noise ratios of around 15 dB could be achieved with the proposed system under a 65 dB babble noise plus directional music noise condition. Subjective intelligibility tests were conducted in the same noise condition. For the words with high familiarity, intelligibility scores increased from 67% to 90% for normal hearing subjects and from 50% to 70% for elderly subjects, when the proposed system was applied.

The results in the present work indicate the potentiality of the proposed system for hearing-impaired people. Future works include implementation of the frequency gain adjustments adapted to the user’s auditory responses, and evaluation by hearing-impaired people. Effect of room reverberation is also a subject for future investigation.

ACKNOWLEDGEMENTS

This research was supported by JST, ERATO, Grant Number JPMJER1401 and MIC/SCOPE. We thank Mika Morita for contributions in conducting the experiments.

REFERENCES

[1] WHO homepage, http://www.who.int/mediacentre/factsheets/fs300/en/ (accessed 1 July 2016).
[2] JapanTrack 2015 report, http://www.hochouki.com/files/JAPAN_Trak_2015_report3.pdf (accessed 1 July 2016).
[3] S. Doclo, S. Gannot, M. Moonen and A. Spriet, “Acoustic beamforming for hearing aid applications,” in Handbook on Array Processing and Sensor Networks, S. Haykin and K. J. Ray Liu, Eds. (Wiley, Hoboken, N.J., 2010), Chap. 9, pp. 269–302.
[4] J. G. Desloge, W. M. Rabinowitz and P. M. Zurek, “Microphone-array hearing aids with binaural output — Part I: Fixed-processing systems,” IEEE Trans. Speech Audio Process., 5, 529–542 (1997).
[5] B. Cornelis, M. Moonen and J. Wouters, “Speech intelligibility improvements with hearing aids using bilateral and binaural adaptive multichannel Wiener filtering based noise reduction,” J. Acoust. Soc. Am., 131, 4743–4755 (2012).
[6] H. Saruwatari, M. Go, R. Okamoto and K. Shikano, “Binaural hearing aid using sound-localization-preserved MMSE STSA estimator with ICA-based noise estimation,” Proc. IWAENC 2010, paper number 1058, 4 pages (2010).
[7] J. I. Marin-Hurtado, D. N. Parikh and D. V. Anderson, “Perceptually inspired noise-reduction method for binaural hearing aids,” IEEE Trans. Audio Speech Lang. Process., 20, 1372–1382 (2012).
[8] A. Bertrand and M. Moonen, “Robust distributed noise reduction in hearing aids with external acoustic sensor nodes,” EURASIP J. Adv. Signal Process., 2009, 1–14 (2009).
[9] T. Noguchi, M. Mizumachi and Y. Nakatoh, “Study of hearing aid system adapted to direction inputted by user,” J. Inst. Ind. Appl. Eng., 3, 72–79 (2015).
[10] B. Widrow and F. L. Luo, “Microphone arrays for hearing aids: An overview,” Speech Commun., 39, 139–146 (2003).
[11] C. Ishi, J. Even and N. Hagita, “Integration of multiple microphone arrays and use of sound reflections for 3D localization of sound sources,” IEICE Trans. Fundam. Electron. Comput. Sci., 97-A, 1867–1874 (2014).
[12] C. Ishi, J. Even and N. Hagita, “Speech activity detection and face orientation estimation using multiple microphone arrays and human position information,” Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS 2015), pp. 5574–5579 (2015).
[13] C. Ishi, C. Liu, J. Even and N. Hagita, “Hearing support system using environment sensor network,” Proc. IROS 2016, pp. 1275–1280 (2016).
[14] C. Liu, C. Ishi and H. Ishiguro, “Bringing the scene back to the tele-operator: Auditory scene manipulation for tele-presence systems,” Proc. ACM/IEEE Int. Conf. Human Robot Interaction, pp. 279–286 (2015).
[15] C. I. Cheng and G. H. Wakefield, “Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space,” J. Acoust. Soc. Am., 49, 231–249 (2001).
[16] D. F. Glas, T. Miyashita, H. Ishiguro and N. Hagita, “Laser tracking of human body motion using adaptive shape modeling,” Proc. IROS 2007, pp. 602–608 (2007).
[17] C. Ishi, O. Chatot, H. Ishiguro and N. Hagita, “Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments,” Proc. IROS 2009, pp. 2027–2032 (2009).
[18] W. G. Gardner and K. D. Martin, “HRTF measurements of a KEMAR,” J. Acoust. Soc. Am., 97, 3907–3908 (1995).
[19] S. Amano, S. Sakamoto, T. Kondo and Y. Suzuki, “Development of familiarity-controlled word lists 2003 (FW03) to assess spoken word intelligibility in Japanese,” Speech Commun., 51, 76–82 (2008).
[20] A. K. Nabelely, M. C. Freyaldenhoven, J. W. Tampas, S. B. Burchfield and R. A. Muenchen, “Acceptable noise level as a predictor of hearing aid use,” J. Am. Acad. Audiol., 17, 626–639 (2006).
[21] J. R. Burwickel, M. F. McKinney and J. Galster, “Acceptable hearing aid throughput delay for listeners with hearing loss under noisy conditions,” Poster at the Annual Sci. and Tech. Conf. of the Am. Auditory Soc. (2016).
[22] J. Even, J. Furrer, L. Y. Morales Saiki, C. T. Ishi and N. Hagita, “Probabilistic 3D mapping of sound-emitting structures based on acoustic ray casting,” IEEE Trans. Robot. (T-RO), 33, 333–345 (2017).