HANDSFREE VOICE INTERFACE FOR HOME NETWORK SERVICE USING A MICROPHONE ARRAY NETWORK

Shimpei Soda, Masahide Nakamura, Shinsuke Matsumoto,
Shintaro Izumi, Hiroshi Kawaguchi, Masahiko Yoshimoto

Graduate School of System Informatics, Kobe University
1-1 Rokkodai, Nada, Kobe, Hyogo, 657-8501 Japan
soda@ws.cs.kobe-u.ac.jp

ABSTRACT

The voice control is a promising user interface for the home network system (HNS). In our previous interface, a user had to be equipped with an actual microphone device, which imposed a burden on the user. This paper presents a hands-free voice interface using a microphone array network. The microphone array network enables voice quality enhancement, as well as sound source localization, by networking multiple microphone arrays. Attaching the arrays to the walls or ceiling, users can input voice operations to the HNS from anywhere in the room, without being aware of the microphone devices. We implement a prototype system with a 16ch microphone array, and evaluate the speech recognition rate and the accuracy of sound source localization in a real home network environment. A hands-free operation service and an automatic speech logging service are implemented.

Index Terms— microphone array network, home network services, voice interface, hands free

1. INTRODUCTION

The *home network system* (HNS) is a core technology of the next-generation smart house, achieving value-added services by networking various household appliances and sensors [1]. In the HNS, a variety of services and appliances are deployed in individual house environment. Therefore, an intuitive and easy-to-learn user interface is required. For this, the *voice interface* is a promising technology, which allows users to operate appliances and services by voice. Since the user can operate a variety of appliances and services by the speech only, it is easy to learn compared to the conventional controllers or panels. We have previously built a mixed-initiative voice interface[2] on the actual HNS.

However, most of the conventional voice interface require to use the close-talking microphone. The users have to be aware of the microphone during the operation, as the microphone should be handed or attached with a head-set. The use of such microphone devices in daily life burdens a significant constraint for the users.

In this paper, we propose a hands-free voice interface using a *microphone array network* [3], which allows users to use the interface without having explicit microphones. In the microphone array network, multiple microphone arrays are collaborated through a network. It can enhance voice quality, estimate a sound location, and separate multiple sound sources using the arrival time differences among microphones[4][5]. By deploying microphone arrays on a wall or ceiling, users can give the voice commands to the HNS from anywhere in the room without regard to the microphone. In this paper, we implement a prototype system using a 16ch microphone array. The speech recognition rate and the accuracy of sound source localization of the prototype are evaluated in the real home network environment. To demonstrate the usefulness of the proposed system, we also implement two practical services: a hands-free operation service and an automatic speech logging service.

2. PRELIMINARIES

2.1. Microphone Array Network

The *microphone array* is a sound collecting device equipped with multiple microphones. Using the difference of arrival time of a sound captured by each microphone, the array can estimate the direction of the sound source and control the directivity. Moreover, by suppressing the effects of reflections and reverberation, the array can separate the noise and extract a particular voice. The signal-to-noise ratio (SNR) can be improved. The performance of the microphone array can be improved significantly with the number of microphones. However, the computational complexity increases polynomially [6] and more energy is required. To satisfy the requirement of ubiquitous sound acquisition, it is necessary to achieve a low-power and efficient sound-processing system.

To cope with the problem, we have proposed to divide the huge array into *sub-arrays* communicating via a network, so called *microphone array network* [3]. The performance can be improved by increasing the sub-arrays. However, the communication between sub-arrays does not increase so much.

978-0-7695-4893-7/12 $26.00 © 2012 IEEE
DOI 10.1109/ICNC.2012.37

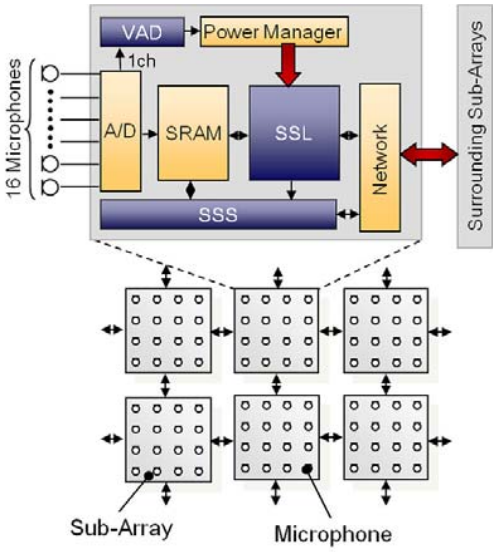

Fig. 1. Microphone array network.

Fig. 1 presents a brief description of the proposed microphone array network and a functional block diagram of a sub-array. In each sub-array, 16ch of microphone inputs are digitized with A/D converters, and stored in SRAM. Each sub-array can perform the following three operations.

Voice Activity Detection(VAD) : detects the presence or absence of speech.

Sound Source Localization(SSL) : estimates the position of the sound source.

Sound Source Separation(SSS) : enhances the quality of sound arriving from a specific location.

Using these operations, each sub-array yields a high SNR audio data. By aggregating these data over the network, the SNR can be improved further. Our latest results cover fundamental studies only, including verification of prototype [4] and complexity reduction of communications [5]. Research of applications and services is our next challenge.

2.2. Home Network System

The *home network system* [1] consists of a variety of household appliances (e.g., room light, television), and sensors (e.g., thermometer, hygrometer). The appliances and sensors are connected via a network. Each device has control API to allow users or external agents to control the device over the network. The HNS is a core technology of the next-generation smart house to provide value-added services. The services include personal home controllers, autonomous home control with contexts like a user's situation and external environment, etc.

In our research group, we have implemented an actual HNS environment, called CS27-HNS. Introducing the concept of service-oriented architecture (SOA) [7], the CS27-HNS integrates heterogeneous and multi-vendor appliances

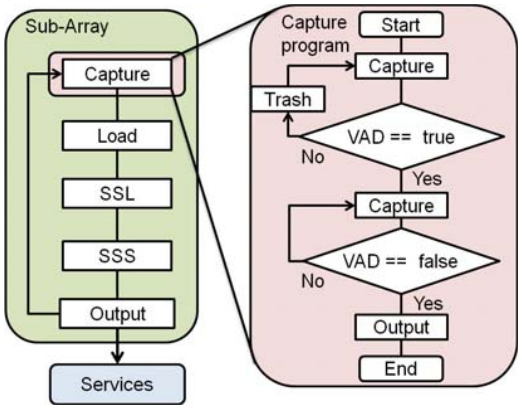

Fig. 2. Processes performed by the proposed system.

by standard Web services. Since the every API can be executed by SOAP or REST Web service protocols, it does not depend on a specific vendor or execution platform.

3. IMPLEMENTING HNS VOICE INTERFACE WITH MICROPHONE ARRAY NETWORK

3.1. System Requirements

We address the three requirements for the target system.

Requirement R1: The system should not burden users. The conventional voice control often required users to speak with a microphone device close to mouth. However, carrying the microphone every time in daily life is quite uncomfortable. Preferably, the voice interfaces for the HNS should be able to be used even without wearing microphone devices.

Requirement R2: The system should tolerate noisy environment. In general, a house is full of various sounds, including TV sounds, air-conditioning, dish washing, etc. Even in such a noisy environment, the system should be able to capture and extract the target sound clearly, by suppressing surrounding noise.

Requirement R3: The system should cover every corner of the room. In daily life, users operate appliances and services in various locations, for instance, on the couch, in front of the door, in the kitchen, etc. Moreover, to implement the *location-aware* services, it is necessary to cover the wide area of the room that any voice may occur.

3.2. Prototype System

To satisfy Requirements R1 to R3, we have implemented a prototype of the voice interface using a single sub-array. The prototype is currenly intended to achieve Requirements R1 and R2, only. Requirement R3 can be achieved by increasing the number of the sub-array, and thus it is beyond this paper.

Fig. 2 shows a flowchart describing the overview of the voice capturing process performed by the prototype system.

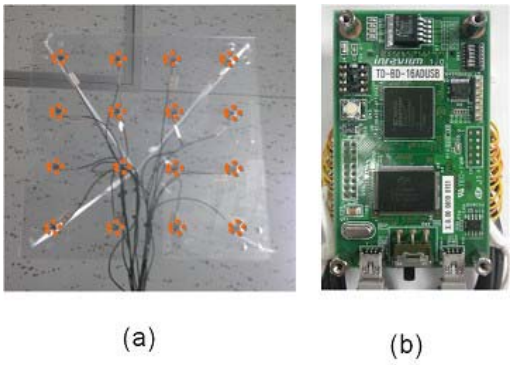

Fig. 3. (a) Sub-array device and (b) capture module.

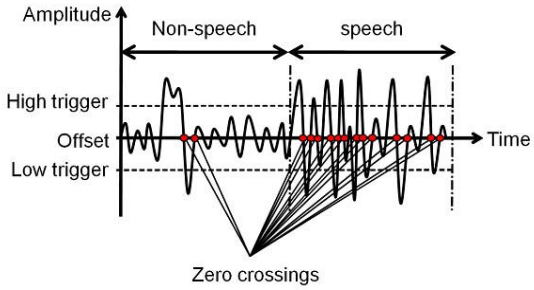

Fig. 4. Example of zero-crossing algorithm. The offset line shows the direct current (DC) component.

In our system, the voice activity detection (VAD) and the capturing program are implemented in C++. The sound source localization (SSL) and the sound source separation (SSS) are implemented in MATLAB.

In the capture program, a voice activity is detected from audio signals collected by 16 microphones in the sub-array. When the voice activity is detected, the system records the voice and outputs each channel. Next, 16ch voice data output from the capture program is loaded on the MATLAB. Then, the SSL estimates the location of the sound source. Based on the estimated sound location, 16ch voice data are aggregated to 1ch by the SSS. Finally, the high-quality 1ch voice data is output to the HNS services.

3.3. Sub-Array / Capture Module

Fig. 3 (a) shows the developed sub-array. The size is a 30 cm square, and 16 microphones (ECM-C10; Sony Corp.) are placed in a grid. The voice data acquired by the microphones is transfered to the PC through the capture module (TD-BD-16USB; Tokyo Electron Ltd.), which is shown in Fig. 3 (b). The capture program executes the VAD to start and stop of the voice recording.

3.4. Voice Activity Detection (VAD)

Our system extensively uses the *zero-crossing algorithm* [8] for the VAD. Fig. 4 depicts the zero-crossing algorithm.

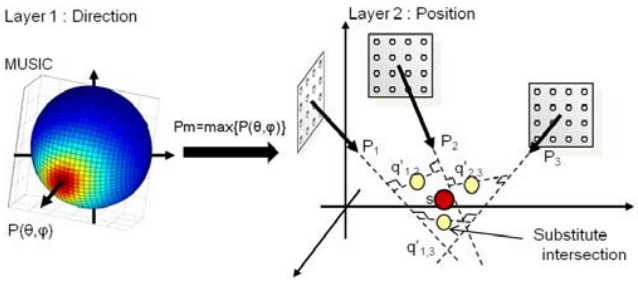

Fig. 5. Three-dimensional sound source localization.

The zero crossing is the first intersection between an input signal and an offset line after the signal crosses a trigger line: the high trigger line or the low trigger line. Between a speech signal and non-speech signal, the appearance ratios of this zero crossing differ. The zero-crossing VAD detects this difference.

We have set the sampling frequency to 1.6 kHz, and the number of bits per sample to 32 bits. Also, we have defined the number of voice samples per frame as 64. A certain number of frames of voice data of all channels are held by the system even while waiting. 1ch of data is used for VAD. A voice recording is started when the system regards that all the frames are in the utterance section. The recording is stopped when a specified number of frames are successively regarded as silent intervals.

3.5. Sound Source Localization (SSL)

In the microphone array network, we divide the sound source localization into two layers: 1) relative direction estimation within a sub-array, and 2) absolute location estimation by exchanging results through the network.

The *MUSIC algorithm*[9] is chosen for sub-array layer estimation because microphones on the sub-array are limited to 16; this algorithm can achieve higher resolution with fewer microphones. To find a relative direction, the sound source probability for $P(\theta, \phi)$ is calculated for each sub-array.

We then localize the absolute sound source location in the network layer. A brief description of this method is presented in Fig. 5 with a three-dimensional coordinate of the sound source. We alternatively adopt the shortest line segment that connects two lines because we can usually find no exact intersection in three-dimensional space. We infer a point that divides the shortest line segment by the ratios of $P(\theta, \phi)$s as an intersection. The sound source is localized by calculating the center of gravity, as well, using the obtained intersections.

3.6. Sound Source Separation (SSS)

Two major approaches used for the SSS are the geometric techniques with position information, and the statistical techniques without position information. The proposed system

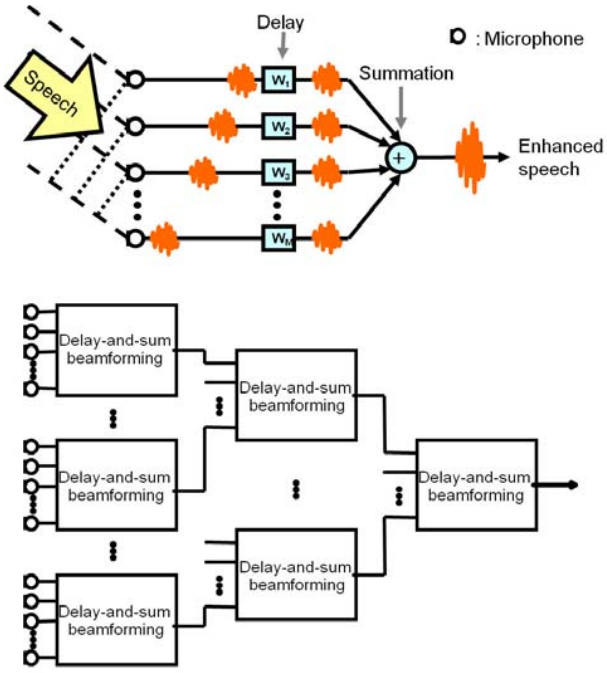

Fig. 6. Delay-and-sum beamforming / distributed processing.

uses one of the former approach, *delay-and-sum beamforming* [10], since the position of sub-array is fixed. This method produces less distortion than statistical techniques; moreover, it requires few computations.

In the delay-and-sum beamforming, multiple signals arriving to microphones with time differences are superposed so that the phase differences are adjusted by delays. As shown in Fig. 6, the phase difference is calculated from estimated sound source location. Thus, only the sound from a specific location is enhanced by the superposition principle. Since the method uses mathematical summation only, we can apply distributed processing using multiple arrays over network.

4. APPLICATION SERVICES

As an application of the proposed system, this section introduces two practical HNS services: (a) *hands-free operation service* and (b) *automatic speech logging service*.

4.1. Hands-free Operation Service

In our CS27-HNS (see Section 2.2), a close-talking microphone has been used for voice operation of HNS appliances. The proposed system enables a *hands-free operation* without such microphone devices, to operate various HNS appliances and services. Fig. 7 shows a brief description of this service.

In the figure, the users are turning on a television and an air conditioner while sitting on a couch. Thus, users can input voice commands from various locations of the room without microphone or controller. Therefore, user's burden is reduced

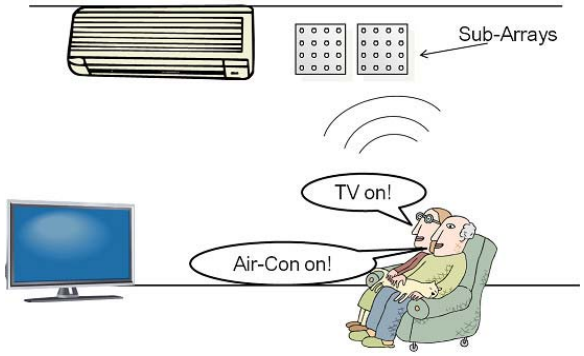

Fig. 7. Hands-free operation service.

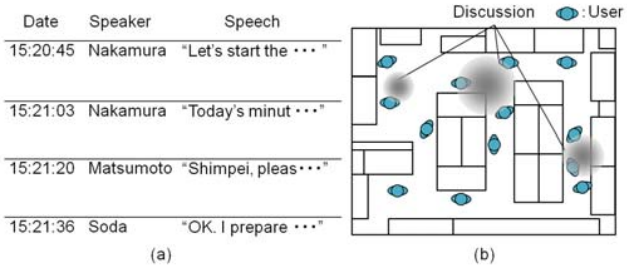

Fig. 8. Automatic speech logging service. (a) Speech log and (b) Speech geographical distribution.

dramatically. Since the voice commands must be delivered correctly in a noisy environment, the key metric for this service is *accuracy of sound recognition*. The microphone array network suppresses the noise by enhancing the voice from the estimated location.

4.2. Automatic Speech Logging Service

Using the SSL feature, the proposed system can associate location information with each voice recorded. The *automatic speech logging service* automatically accumulates the speech data with date, time, and location information. The data can be used as *lifelog* within a house (or an office), with which users can review what speech occurred when and where.

Fig. 8 shows a brief description of the service. Cooperating with a voice recognition module, it is possible to perform the automatic dictation of meeting, as well as a protocol analysis of interview. By plotting sound sources on a map, distribution of speech can be visualized, showing where users often speak in the room. An interesting challenge is to evaluate the motivation of participants in a meeting by speech contents and the number of utterances. The VAD feature allows the service to record the voice only while somebody speak, which significantly reduces the size of storage. The key metric for this service is the *accuracy of the SSL*.

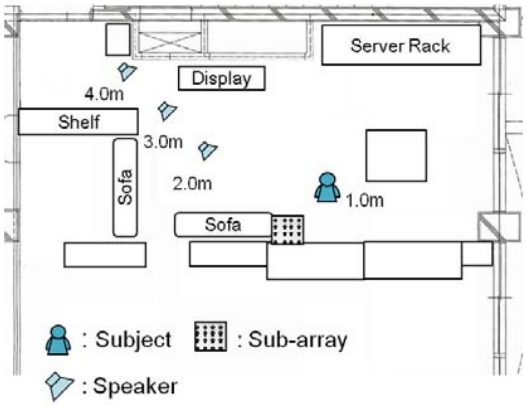

Fig. 9. Position of subjects and devices.

Table 1. Experiment 1: recognition ratio of each subject.

Subject No.	1	2	3	4	5
Age	twenties	twenties	twenties	twenties	forties
Sex	Male	Male	Male	Male	Male
Recognition rate (%)	90	80	82	94	90

5. EXPERIMANTAL EVALUATION

To see the feasibility of the HNS services, we evaluate the speech recognition rate and the accuracy of sound source localization in this section.

5.1. Speech Recognition Rate

The hands-free operation service requires high recognition rate of user's voice command. To evaluate the recognition rate, we deployed the prototype in the CS27-HNS, and asked each subject to speak operation commands of the CS27-HNS. Fig. 9 shows the layout of our experimental room. We conducted two kinds of experiments:

Experiment 1: We measure the variance of the recognition rate for different users. Each of five subjects speaks 50 voice commands at the position of 1.0m from the sub-array.

Experiment 2: We measure the variance of the recognition rate by different distance. Each of three speakers placed in different locations (see Fig. 9) plays 50 voice commands recorded in Experiment 1. Positions of the speakers are at 2.0m, 3.0m, and 4.0m away from the sub-array, respectively.

Table 1 shows the result of Experiment 1. Each row represents a subject number, age, gender, and recognition rate. Despite of difference in pronunciation of each subject, the prototype achieved quite high recognition rate from 80% to 94%. The recognition rates of subject 2 and subject 3 were relatively low. This was because for some commands, the subject did not speak the first or last letter, clearly.

Fig. 10 shows the result of Experiment 2. The horizontal axis represents the distance between the speakers and sub-

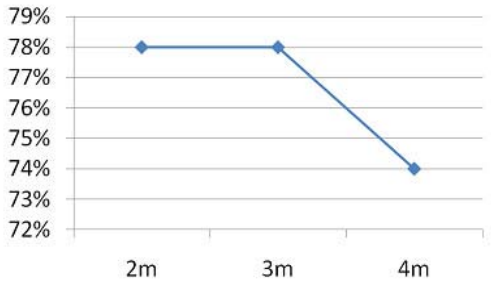

Fig. 10. Experiment 2: recognition ratio at each distance.

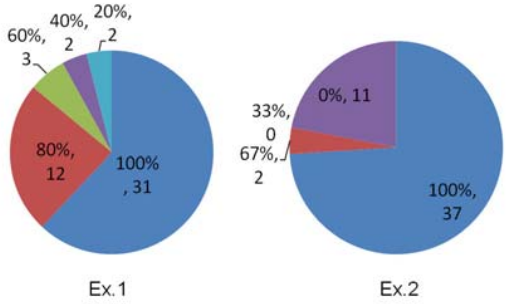

Fig. 11. Percentage of recognition rate of each word.

array. The vertical axis is the recognition rate. As the distance increases, the recognition rate declines due to playback noises from the speaker. Even so, 74% recognition rate was achieved at the position of 4.0m from the sub-array. From these results, it seems that the proposed system is sufficiently feasible to implement the hands-free operation service.

Fig. 11 shows the percentage of recognition rate for each word in the two experiments. A number besides a percentage represents the number of commands recognized with that recognition rate. In Experiment 1, more than 90% of the all commands were recognized with more than 80% recognition rate. In Experiment 2, however, 11 commands had not been understood at all, regardless of the distance.

In order to increase the recognition rate, it is necessary to encourage users to speak clearly and loudly. Also, we need to enhance the noise reduction by the sound source separation. We deployed only a single sub-array in this paper. However, we will increase the number of sub-array to expand coverage and improve the performance of the SSS.

5.2. Acurracy of SSL

The automatic speech logging service requires high accuracy of sound the source localization (SSL). For this, we recorded a regular meeting in our laboratory using the prototype system. The recording time was about 16 minutes. When the system detects a voice activity, the system records the speech data, time stamp, and the direction of arrival. We evaluated manually whether the estimated direction was consistent with the speaker's position. The number of participants was eight,

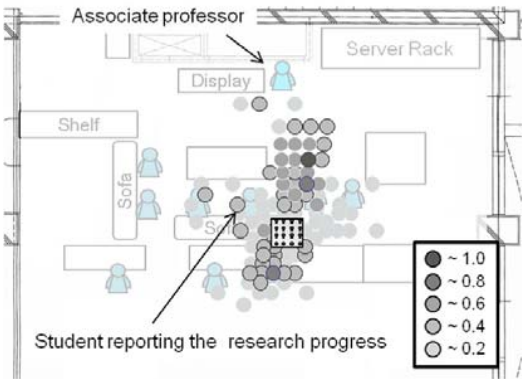

Fig. 12. Distributuion of positions of utterances.

consisting of an associate professor, an assistant professor, and six students. The layout of the meeting room is the same as the one shown in Fig. 9.

Fig. 12 shows the distribution of speeches in the meeting, estimated by the prototype system. A circle represents coordinates of the speech estimated by the SSL. The color depth of the circle represents the frequency of the speech yielded in that position. The number of speeches is normalized based on a place that has the maximum frequency.

Those who spoke most frequently in the meeting were the the associate professor and a student reporting the research progress (see Fig. 12). Therefore, it can be seen that the distribution is skewed in the direction of the left side of the display. However, the position of associate professor did not match the estimated position, because associate professor was outside the coverage area of the sub-array. To expand the coverage and improve the accuracy of SSL, we need to deploy more sub-arrays, which is left for our future work.

6. CONCLUSION

In this paper, we have proposed to use a microphone array network to achieve practical hands-free voice interface for the home network system (HNS). We have implemented a prototype system using a 16ch sub-array, and evaluated it with an actual HNS. As a result, the prototype system achieved high recognition rate of 80% to 94% at close range, and 74% at distant range of 4.0m.

Our future work is to deploy more sub-arrays, in order to expand the coverage and improving the accuracy of the sound source localization and voice activity detection. We also study other HNS services using the microphone array network, and perform more experiments to show the effectiveness.

7. ACKNOWLEDGMENTS

This research was partially supported by the Semiconductor Technology Academic Research Center (STARC), the Japan Ministry of Education, Science, Sports, and Culture [Grant-in-Aid for Scientific Research (C) (No.24500079), Scientific Research (B) (No.23300009)], and Kansai Research Foundation for technology promotion.

8. REFERENCES

[1] M.Nakamura, A.Tanaka, H.Igaki, H.Tamada, and K.Matsumoto, "Constructing home network systems and integrated services using legacy home appliances and web services," *International Journal of Web Services Research*, vol. 5, no. 1, pp. 82–98, 2008.

[2] M. Nakamura N. Matsubara, S. Matsumoto, "Characterizing user habituation in interactive voice interface - experience study on home network system," in *The 13th International Conference on Information Integration and Web-based Applications & Services (iiWAS)*, 2011, vol. 109, pp. 61–66.

[3] T. Takagi, H. Noguchi, K. Kugata, M. Yoshimoto, and H. Kawaguchi, "Microphone array network for ubiquitous sound acquisition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 1474–1477.

[4] H. Noguchi M. Yoshimoto K. Kugata, T. Takagi and H. Kawaguchi, "Intelligent ubiquitous sensor network for sound acquisition," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2010, pp. 585–588.

[5] S. Izumi, H. Noguchi, T. Takagi, K. Kugata, S. Soda andM. Yoshimoto, and H. Kawaguchi, "Data aggregation protocol for multiple sound sources acquisition with microphone array network," in *20th International Conference on Computer Communications and Networks (ICCCN)*, 2011, pp. 1–6.

[6] Cairns Australia and James Glass, "Loud: A 1020-node microphone array and acoustic," 2007.

[7] M.P.Papazoglou and D.Georgakopoulos, "Service-oriented computing," *Communication of the ACM*, vol. 46, no. 10, pp. 25–28, 2003.

[8] M. M. Sondhi J. Benesty and Y. Huang, *Springer Handbook of Speech Processing*, Springer-Verlag, 2008.

[9] R. Schmidt, "Multiple emitter location and signal parameter estimation," *Antennas and Propagation, IEEE Transactions on*, vol. 34, pp. 276–280, 1986.

[10] K. Buckley Van Veen, "Beamforming: a versatile approach to spatial filtering," *ASSP Magazine, IEEE*, vol. 5, pp. 4–24, 1988.