社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

Introducing Multiple Microphone Arrays for Enhancing Smart Home Voice Control

Shimpei SODA[†], Masahide NAKAMURA[†], Shinsuke MATSUMOTO[†],
Shintaro IZUMI[†], Hiroshi KAWAGUCHI[†], and Masahiko YOSHIMOTO[†]

[†] Graduate School of System Informatics, Kobe University
1-1 Rokkodai, Nada, Kobe, Hyogo, 657-8501 Japan

Abstract We have previously developed a voice control system for a home network system (HNS), using a microphone array technology. Although the microphone array achieved a convenient hands-free controller, a single array had limitations on coverage of sound collection and speech recognition rate. In this paper, we try to overcome the limitations by increasing the number of the microphone arrays. Specifically, we construct a *microphone array network* using four separate arrays, and enhance algorithms of sound source localization (SSL) and sound source separation (SSS) on the network. We also conduct an experimental evaluation, where precision of SSL and speech recognition rate are evaluated in a real HNS test-bed. As a result, it is shown that the usage of multiple arrays significantly improves the coverage and speech recognition ratio, compared with the previous system.

Key words microphone array network, multiple microphone arrays, smart home, 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.

The *voice control* is a promising user interface for the HNS, 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. However, most conventional systems require users to use explicit microphone devices, which is a burden on daily life in the house. To cope with the problem, we are studying a hands-free voice interface using a *microphone array technology* [2]. A microphone array, comprised of multiple microphones in a grid form, is a device for collecting high-quality sound within indoor space. Using time differences of sound arriving to different microphones, it can enhance voice quality, estimate a sound location, and separate multiple sound sources [3] [4]. By installing the microphone arrays on a wall or ceiling, users can give the voice commands to the HNS from anywhere in a room without realizing explicit microphone devices. In our previous work [5], we have implemented a prototype system using a 16ch *single* sub-array. However, the single array could not achieve sufficient performance for practical use, specifically, with respect to the coverage of sound collection and speech recognition ratio.

In this paper, we try to overcome the limitations by increasing the number of the microphone arrays. The previous single array is now extended to a *microphone array network*, comprised of four separate 4ch arrays. Algorithms of sound source localization (SSL) and sound source separation (SSS) are also revised to adapt to the multiple arrays. Finally, we conduct an experimental evaluation of the developed system within a real HNS test-bed. The result shows that the coverage of sound collection is significantly expanded, and that the speech recognition rate is improved more than 70% within 5.0m in radius from the microphone arrays.

2. Previous Work

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

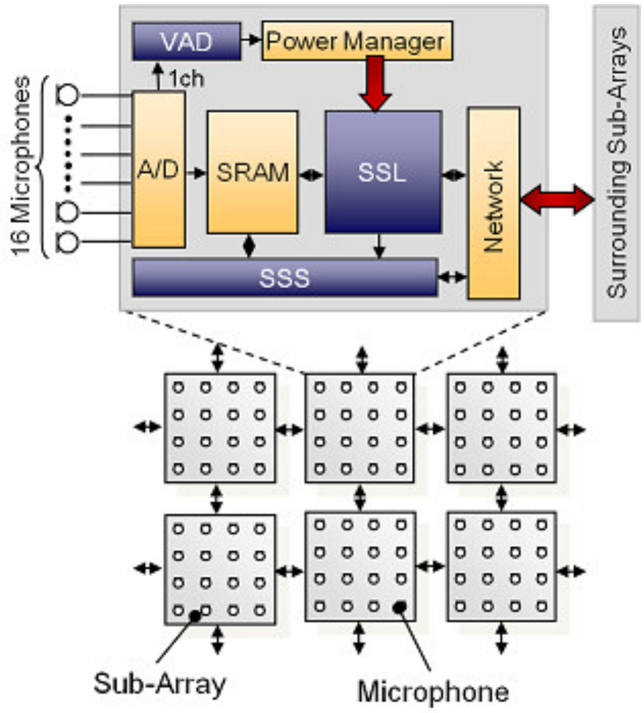


Fig. 1 Microphone array network.

(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* [2]. The performance can be improved by increasing the sub-arrays. However, the communication between sub-arrays does not increase so much. 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. We have been studying the microphone array network from the fundamental and theoretical aspect. The results include verification of prototype [3] and complexity reduction of communications [4]. Design and implementation of *practical systems* using the microphone array network is our important challenge. The application to the HNS, presented in this paper, is one of such practical systems.

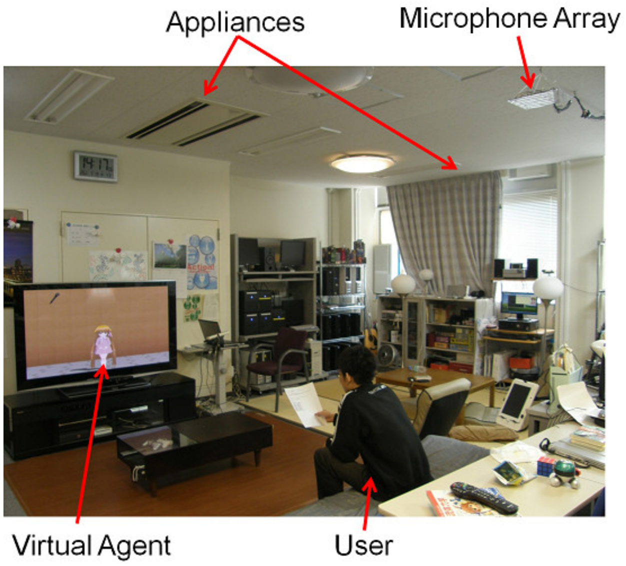


Fig. 2 Hands free voice interface using virtual agent.

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 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. Fig. 2 shows the experimental room of CS27-HNS.

2.3 Hands Free Voice Interface

Since a variety of appliances and services are deployed in the HNS, intuitive and easy-to-learn human interface to control the HNS is required. The *voice interface* is a promising technology to implement a universal controller of the HNS, since it can abstract heterogeneous operations in terms of speech. Most conventional voice interfaces require using close-talking microphones (e.g., ones with headsets or smartphones). However, carrying always such microphone devices everywhere in the house burdens significant constraint in the daily life.

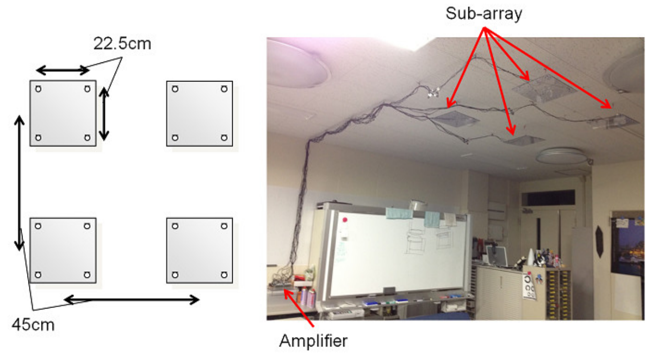


Fig. 3 Sub-arrays installed in ceiling of CS27-HNS.

To cope with the problem, we are developing a *hands-free voice controller with a microphone array* [5], built in a ceiling of CS27-HNS. The system is intended to allow users to speak from everywhere without being aware of microphones, and to achieve good quality of voice sampling in noisy environment. As shown in Fig. 2, the previous prototype used a *single* microphone array. In addition, we are also employing the *virtual agent* technology [8] [9], which can introduce affinity and humanity in spoken dialog systems. By integrating a virtual agent with our hands-free controller, we expect that a user can enjoy operating the HNS through more natural conversations with the agent. In Fig. 2, a user is talking to an agent displayed on a TV, in order to operate appliances.

2.4 Limitations of Previous Prototype

In our preliminary evaluation, the previous prototype had the following limitations for practical use.

- The speech recognition rate was about 60%, which often mis-recognized appliance operations.
- The coverage of sound source localization (SSL) was only 1.0 m in radius from the microphone array.
- The system could not tolerate noisy environment.

The major cause of the limitations is that the prototype had a single microphone array only. By increasing the number of arrays, we could expect to overcome the limitations.

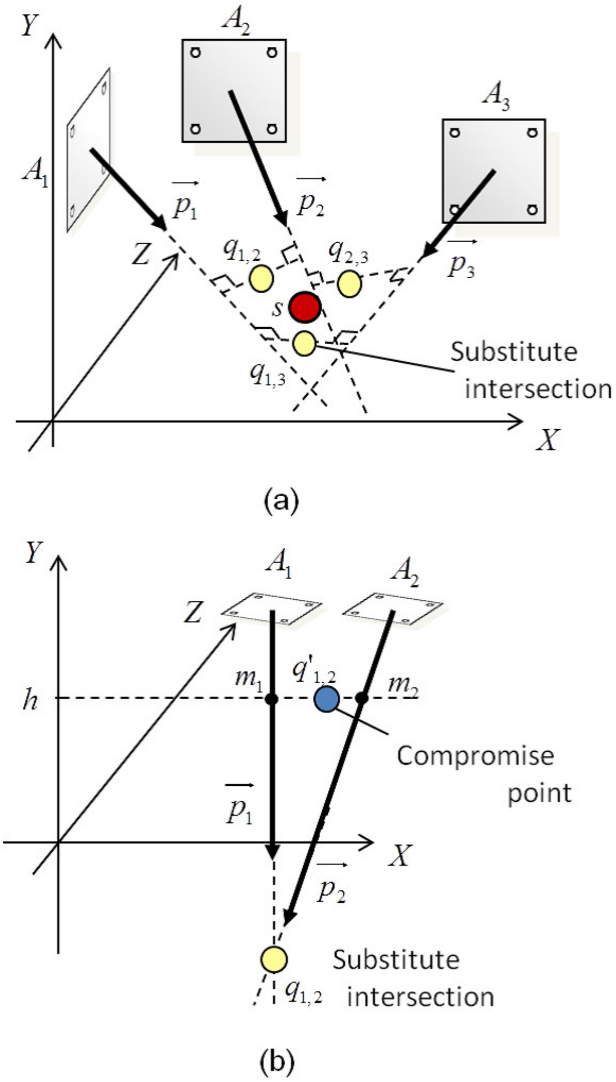
3. Extension to Multiple Arrays

The goal of this paper is to deploy extra arrays to cope with the above limitations. For this, we consider how to place the arrays and revise the algorithm of SSL to adapt to the multiple arrays. We then evaluate again the precision of SSL and speech recognition rate, with the multiple arrays.

3.1 Placement of Sub-Arrays

Fig. 3 shows the placement of microphones and sub-arrays in the proposed system. Each sub-array has four microphones in the each corner of a square acrylic plate. The acrylic plate is a square 30 cm and the interval between a pair of microphones is 22.5 cm.

As shown in Fig. 3, the four sub-arrays are arranged in square in the ceiling. In our preliminary study [5], it was shown that the distance between a pair of sub-arrays should be wide to improve the sound source localization (i.e., coverage of the system), while the distance should be short to improve the sound source separation (i.e., quality of sound). We embed hooks in the ceiling so that we can suspend the sub-arrays with 3 different distance configurations: 45 cm, 90 cm and 135 cm. In this paper, we take the medium configuration, i.e. 90cm, to evaluate the whole system.



(a)

(b)

Fig. 4 How to calculate compromise point.

3.2 Sound Source Localization (SSL) with Multiple Arrays

To achieve SSL with the four sub-arrays, we choose *MUSIC algorithm* [10]. This algorithm can achieve high resolution of sound localization with a relatively few microphones. The algorithm first estimates, for each sub-array, a relative direction of a sound source by calculating sound source probability $P(\theta, \phi)$.

The algorithm then localizes the absolute sound source location by obtaining intersection of the estimated directions.

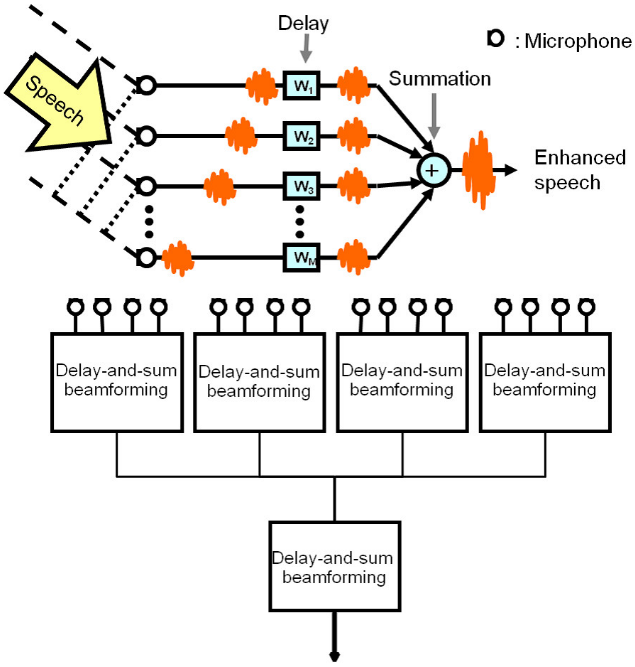

Delay
: Microphone
Summation
Speech
W_1
W_2
W_3
W_M
Enhanced speech
Delay-and-sum beamforming
Delay-and-sum beamforming
Delay-and-sum beamforming
Delay-and-sum beamforming
Delay-and-sum beamforming

Fig. 5 Delay-and-sum beamforming / distributed processing.

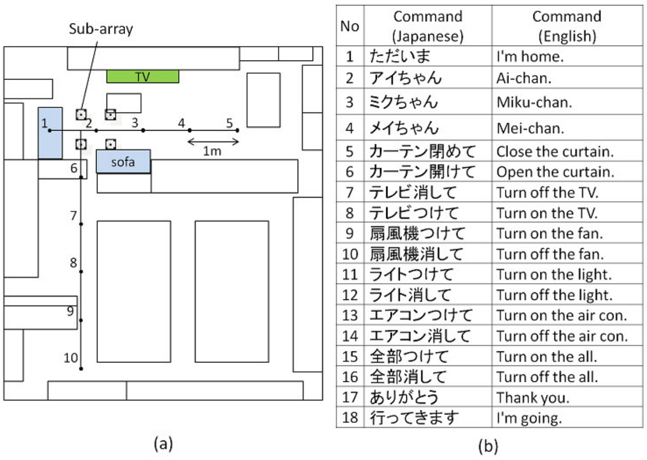
A brief description is presented in Fig. 4(a). In a three-dimensional space, we do not always obtain exact intersection. Hence, we alternatively adopt the shortest line segment that connects two vectors p_i and p_j. We infer a point q_{ij} that divides the shortest line segment by ratios of $P(\theta, \phi)$'s. The sound source s is virtually determined as a center of gravity from the obtained intersections.

In a real environment, however, the virtual intersection q sometimes points a physically improbable position (e.g. under the floor or above the ceiling). In this case, we calculate a compromised point to determine the final location of the sound source. Fig. 4(b) shows how to obtain the compromised point q'. When q is physically improbable, a points m_1 and m_2 are derived from p_1 and p_2 as the intersections of a pre-determined height h. In the proposed system, h is 160cm which is close to the average height of a mouth of a user. The compromised point q' is determined on the straight line m_1m_2, so that q' divides m_1m_2 by the ratios of p_1 and p_2.

3.3 Sound Source Separation (SSS) with Multiple Arrays

The proposed system uses one of the former approach, *delay-and-sum beamforming*[11], 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. 5, 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.



No	Command (Japanese)	Command (English)
1	ただいま	I'm home.
2	アイちゃん	Ai-chan.
3	ミクちゃん	Miku-chan.
4	メイちゃん	Mei-chan.
5	カーテン閉めて	Close the curtain.
6	カーテン開けて	Open the curtain.
7	テレビ消して	Turn off the TV.
8	テレビつけて	Turn on the TV.
9	扇風機つけて	Turn on the fan.
10	扇風機消して	Turn off the fan.
11	ライトつけて	Turn on the light.
12	ライト消して	Turn off the light.
13	エアコンつけて	Turn on the air con.
14	エアコン消して	Turn off the air con.
15	全部つけて	Turn on the all.
16	全部消して	Turn off the all.
17	ありがとう	Thank you.
18	行ってきます	I'm going.

Fig. 6 (a) Experiment environment and sound source positions. (b) List of available commands.

4. Evaluation

4.1 Overview of Experiment

We have integrated the proposed microphone array network to CS27-HNS hands-free voice interface (see Section 2.3). We have conducted an experiment to evaluate accuracy of the SSL and speech recognition rate. Five subjects participated in the experiment, each of the subjects speaks 18 voice commands of operating CS27-HNS. Fig. 6(a) shows the experimental environment. Evaluation was performed at 10 different locations in the room shown in Fig. 6(a). For each location, we measure the recognition rate of voice commands and the error of SSL.

At the locations from no.1 to no.5, we compare two environment setting; one is noisy and the other is calm, to see the tolerance of noise. In the noisy environment, TV sound is used as the noise source. Fig. 6(b) enumerates the voice commands that subjects speak in the evaluation. The voice commands involves the ones that starts or terminates the system, and the ones that turns on / off the appliances.

4.2 Speech Recognition Ratio

Fig. 7 shows the recognition ratio in each location. The horizontal axis represents the location number illustrated in Fig. 6(a). The vertical axis is the average recognition rate of the five subjects. In the locations from no.1 to no.5, the recognition ratios in the noisy environment are also shown. The graph shows that the recognition ratio was about 80% in the close range. Even within 5.0m in radius, over 70% recognition rate was achieved. In the noisy environment, recognition rate slightly decreased to 5% to 10%.

Fig. 8 shows the coverage of the proposed system based on the recognition rate. The area where the recognition rate

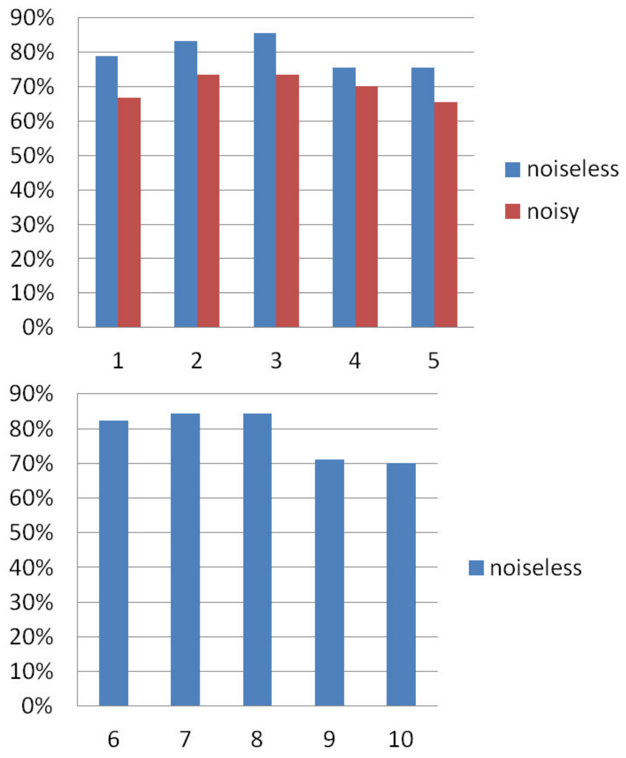


Fig. 7 Recognition rate in each location.

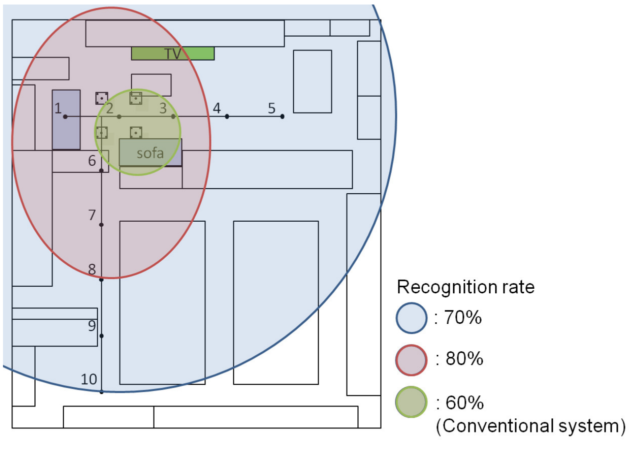


Fig. 8 Coverage of proposed system based on recognition rate.

is over 70% is represented by the outer circle, and the one over 80% is drawn in the second circle.

The innermost circle shows the coverage of the previous prototype with a single array, in which the recognition ratio is about 60%. It can be seen from Fig. 8 that the recognition rate and the coverage has been expanded dramatically by the increase of the number of microphone arrays.

4.3 Accuracy of SSL

Fig. 9 shows the absolute error of sound source localization for each location. The vertical axis is the average error of five subjects. Here the error means a three-dimensional norm between estimated position and actual position of the sound source.

Fig. 10 shows the coverage of the proposed system based

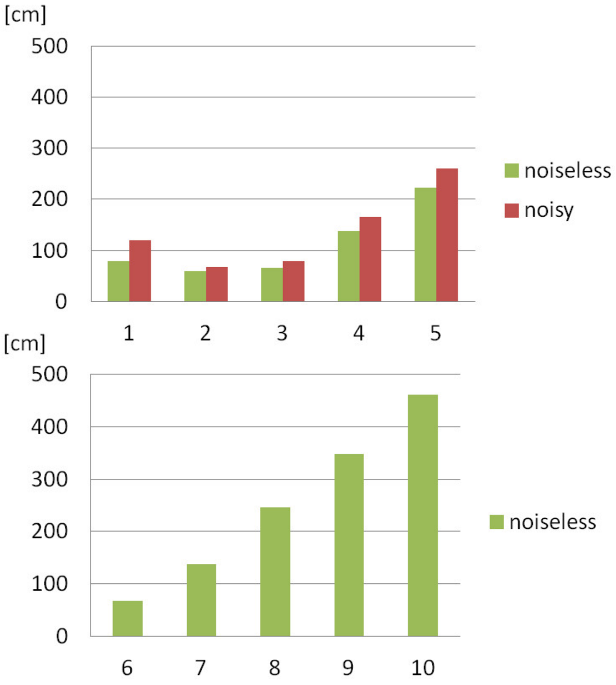


Fig. 9 SSL error in each location.

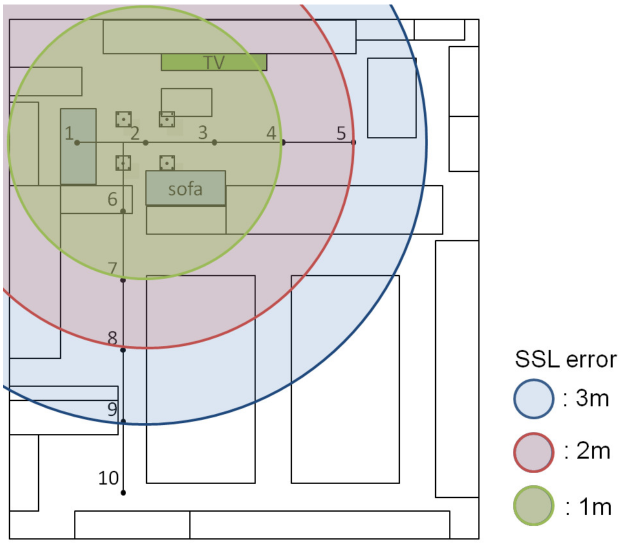


Fig. 10 Coverage of proposed system based on SSL error.

on the SSL error. In the locations no.1 - no.4 and no.6 - no.7, the error is around 1 m. These locations are within 2 m in radius around the sub-array, as shown in the innermost circle in Fig. 10. The error is more than 2m in the location no.8 - no.10, as the distance from the sub-array becomes larger. In Fig. 10, the second circle indicates the coverage where the error is within 2m. The outer circle also indicates the area with 3m error.

In the noisy environment, the error is slightly increased from 8cm to 40cm. This means that the interference of the noise to SSL in closer range was relatively low.

In summary, the following facts were shown in the experiment. When applying to the HNS service that requires high recognition ratio, the proposed system with four sub-array can cover a range of 5m as shown in Fig. 8. As for the services that requires accurate sound source localization (e.g., location-aware voice control), the coverage is around 2m.

5. Related Work

Voice interface with a microphone array is also useful in noisy environment such as outside of building. Oh et al. have proposed a hands-free voice communication system with a microphone array for use in an automobile environment [12]. They have aimed to realize a reliable speech recognition in noisy automobile environment for digital cellular phone application. This study has common purpose with our study that hands free operation for practical applications. Our system should obtain more reliable for noisy environment by introducing their system.

European Media Laboratory has proposed a smart home voice controller using a mobile phone [13]. In this system, a mobile device is used as a close-talking microphone and voice recognition module. Therefore, their whole system is implemented physically-compact compared with common microphone array device including our developed system. Their "compact and mobile" system and our "ubiquitous and mounted" system should be used for different purposes. Because microphone array device included in our proposed system is wrapped as a service, we can easily apply the mobile phone as voice recognition module.

6. Conclusion

In this paper, we developed a hand-free voice control for smart houses using the microphone array technology. To improve the recognition rate and coverage limitations of the previous prototype, we have increased the number of sub-array to four. The algorithms of sound source localization and sound source separation were also revised to adapt multiple sub-arrays. The experimental evaluation in an actual HNS environment showed that the proposed system could significantly improve the coverage and the recognition rate.

Our future works include evaluation of voice activity detection and sound source separation. Also, we compare the performance by different placement configurations of the sub-arrays.

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.

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

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

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

[5] S. Soda, S. Matsumoto, M. Nakamura, S. Izumi, H. Kawaguchi, and M. Yoshimoto, "Handsfree voice interface for home network service using a microphone array network," Techinical Report of IEICE, pp.73–78, March 2012. (in Japanese).

[6] C. Australia and J. 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] Ochs, Magalie, Pelachaud, Catherine, Sadek, and David, "An empathic virtual dialog agent to improve human-machine interaction," Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems - Volume 1, pp.89–96, 2008.

[9] J. Cassell, "Embodied conversational interface agents," Communications of the ACM, vol.43, no.4, pp.70–78, 2000.

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

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

[12] S. Oh, V. Viswanathan, and P. Papamichalis, "Hands-free voice communication in an automobile with a microphone array," Proceedings of the 1992 IEEE international conference on Acoustics, speech and signal processing - Volume 1, ICASSP'92, Washington, DC, USA, pp.281–284, IEEE Computer Society, 1992.

[13] J. Ivanecky, S. Mehlhase, and M. Mieskes, "An intelligent house control using speech recognition with integrated localization," Ambient Assisted Living: 4. AAL-Kongress 2011 Berlin, Germany.