Kobe University Repository : Kernel

PDF issue: 2025-12-05

Data-Intensive Sound Acquisition System with Large-scale Microphone Array

Noguchi, Hiroki ; Takagi, Tomoya ; Kugata, Koji ; Izumi, Shintaro ; Yoshimoto, Masahiko ; Kawaguchi, Hiroshi

(Citation)
Journal of Information Processing,19:129-140

(Issue Date)
2011-03

(Resource Type)
journal article

(Version)
Version of Record

(Rights)
©2011 Information Processing Society of Japan.
利用は著作権の範囲内に限られる

(URL)
https://hdl.handle.net/20.500.14094/90002987

Regular Paper

Data-Intensive Sound Acquisition System with Large-scale Microphone Array

Hiroki Noguchi,[†1] Tomoya Takagi,[†1] Koji Kugata,[†1]
Shintaro Izumi,[†1] Masahiko Yoshimoto[†1]
and Hiroshi Kawaguchi[†1]

We propose a microphone array network that realizes ubiquitous sound acquisition. Several nodes with 16 microphones are connected to form a novel huge sound acquisition system, which carries out voice activity detection (VAD), sound source localization, and sound enhancement. The three operations are distributed among nodes. Using the distributed network, we produce a low-traffic data-intensive array network. To manage node power consumption, VAD is implemented. The system uses little power when speech is not active. For sound localization, a network-connected multiple signal classification (MUSIC) algorithm is used. The experimental result of the sound-source enhancement shows a signal-noise ratio (SNR) improvement of 7.75 dB using 112 microphones. Network traffic is reduced by 99.11% when using 1,024 microphones.

1. Introduction

In recent years, digital human interfaces have been developed for living spaces, medical centers, robotics, and automobiles. Future applications will enable one person to control thousands of microprocessors without being conscious of their existence. Some face and speech recognition systems are in practical use, but most systems operate only in constrained environments or strictly defined installation conditions, with parameters such as an angle or distance to a device. Users must confront a camera in a face recognition system; a microphone must be near a mouth in a speech recognition system. For most people, these constraints are inconvenient for daily life. Therefore, various intelligent ubiquitous sensor systems have been developed as new human interfaces [1]. In future applications, numerous cameras and microphones will be emplaced on walls and roofs of living spaces. They will obtain visual information and speech data automatically and support absolutely hands-free systems. As described herein, we specifically examine speech signal processing as a ubiquitous sensor system because a speech interface is a fundamental mode of human communication; moreover, a speech interface has a much broader range of applications.

Recent improvements in information processing technology have produced real-time sound-processing systems using microphone arrays [2]. One application is a meeting system with a 128-ch square microphone array [3], which captures speech data from every microphone: The microphone array processes signal recordings and also performs noise reduction, sound-source separation, speech recognition, speaker identification, and other tasks. A microphone array can localize sound sources and separate multiple sources using spatial information of the acquired sounds. Huge microphone arrays have been widely investigated: arrays have been built at Tokyo University of Science (128 ch) [3], the University of Electro-Communication (156 ch) [4], Brown University and Rutgers University (512 ch) [5], and the Massachusetts Institute of Technology (1,020 ch) [6]. Nevertheless, their practical use is confounded by difficult problems of increasing computation, power consumption, and network cost, particularly in terms of sound-data acquisition. The salient difficulty of conventional microphone array systems is that all microphones are connected to a single base station (high-performance sound server) with large-scale multi-channel sound recording devices. In conventional systems, concentrative connection of numerous microphones engenders heavy network traffic. The computational effort increases exponentially as the number of microphones increases. If more than 1,000 microphones are used to collect the data, then the signal-noise ratio (SNR) can be improved remarkably [6], but the network traffic and computational amounts skyrocket, effectively prohibiting such systems. To reduce the increased network traffic and computational power of a microphone array system and to satisfy recent demands for ubiquitous sound acquisition, it is necessary to realize a large sound-processing system covering a wide-ranging human environment at low power.

To implement a microphone array as a realistic ubiquitous sound acquisition system with scalability, we propose the division of the huge array into sub-

[†1] Graduate School of Engineering, Kobe University

© 2011 Information Processing Society of Japan

arrays to produce a multi-hop network: an intelligent ubiquitous sensor network (IUSN) [7)–10)]. The sub-array nodes with some microphones can be set up on the walls and ceiling of a room. Reducing the amount of transmission can be accomplished after introducing multi-hop networking. Each relay node on a routing path must store all temporal multi-channel sound data that the node receives, but not send the data. This function demands a large buffer memory and large total power dissipation in the system. Therefore, some breakthrough network solution is necessary to reduce the network traffic. Herein, we describe how our IUSN solves the problems described above. Multi-hop networking, specific data aggregation, and distributed processing are novel concepts unlike conventional microphone array systems. The performance can be improved easily by increasing the node number, but communication among nodes does not increase much in our system.

The distributed sound acquisition system is presented in Section 2. The low-power technique with VAD hardware module is explained in Section 3. Sections 4 and 5 present a discussion of the performance and accuracy of the proposed data acquisition scheme, which is based on sound source localization and sound source separation, using measured data. Section 6 denotes our future works, and Section 7 concludes this paper.

2. Intelligent Ubiquitous Sensor Network and Its Node

This section presents a description of an implementation of the proposed perfect aggregation scheme to a microphone array system.

Figure 1 presents a brief description of the proposed IUSN and a functional block diagram of a sub-array node. Sixteen-microphone inputs are digitized using A/D converters; the sound information is stored in SRAM. Then, the information is used for sound source localization and sound source separation. The sound-processing unit including them is activated by the power manager and voice activity detection (VAD) module to conserve power: The sound processing unit is turned off if no sound exists around the microphone array. Power management is fundamentally necessary because enormous microphones waste much power when they are not in use. In our VAD, the sampling frequency can be reduced to 2 kHz and the number of bits per sample can be set to 10 bits. These values are sufficient to detect human speech, in which case only 3.49 μW is dissipated on a 0.18-μm CMOS process [7)]. By separating the low-power VAD module from the sound processing unit, it can turn off the sound processing unit using the power manager. A single microphone is sufficient to detect a signal. The remaining 15 microphones are turned off as well. Furthermore, not all VAD modules in all nodes need operate. The VAD modules are merely activated in a limited number of nodes in the system.

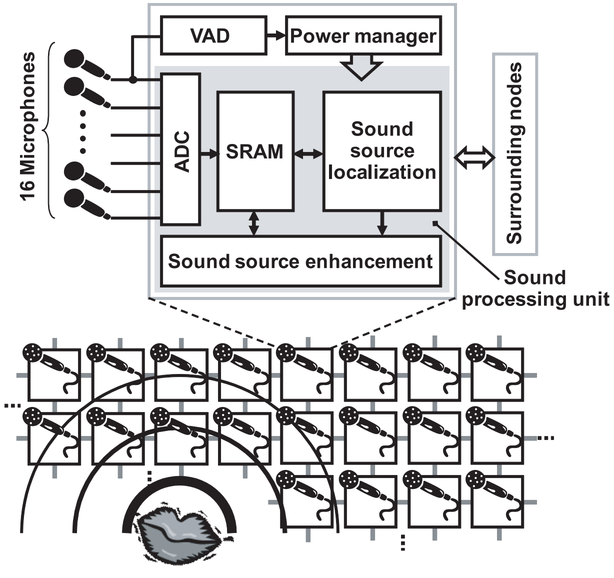


Fig. 1 Intelligent ubiquitous sensor network (IUSN) and block diagram of a sub-array node.

Figure 2 portrays a flow chart of our system. The salient features of the system are: 1) low-power voice activity detection to activate the entire node, 2) sound-source localization to locate sound sources, and 3) sound-source enhancement to reduce the sound noise level. The sub-array nodes are connected to support their mutual communication. Therefore, the sound gained by each node can be gathered to improve the sound source's SNR further. The system constitutes a huge microphone array through interaction with surrounding nodes. Therefore, computations can be distributed among nodes. The system

© 2011 Information Processing Society of Japan

has scalability in terms of the number of microphones. Each node preprocesses acquired sound data. Therefore, only compressed data—localized and enhanced sound—are communicated.

We use low-power zero-crossing VAD, as described in Section 3. Sections 4 and 5 present discussion of the performances and accuracies of the sound-source localization and sound-source enhancement in our system using measured data. For the system, gathering and processing localization data are important to improve the localization accuracy. Distributed localization data obtained using the MUSIC algorithm can be processed by a communication network in our system. Regarding sound-source enhancement, we use basic delay-and-sum beamforming both within a node and among nodes[11]. Therefore, the time accuracy between nodes strongly affects the final SNR of the sound source collected with the network.

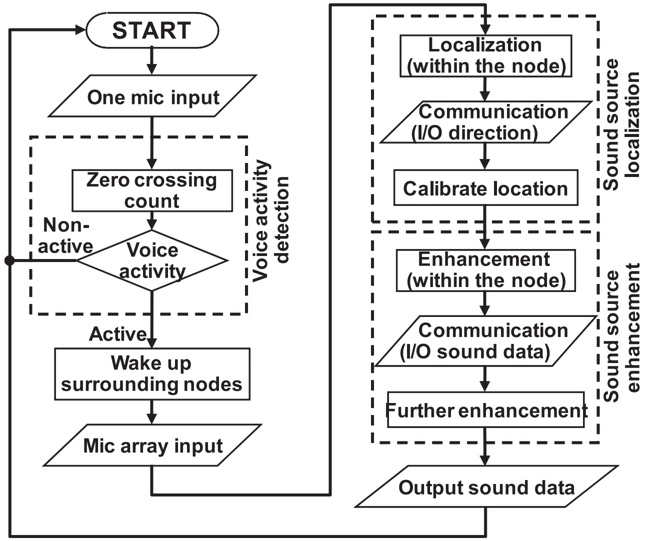


Fig. 2 Flow chart of intelligent ubiquitous sensor nodes.

3. Voice Activity Detection

A microphone array network consists of numerous microphones, whose power consumption would easily become prohibitive. Our intelligent ubiquitous sensor node must therefore operate with a limited energy source to save power to the greatest extent possible. Sound processing that conserves power is effective because the sound processing unit and microphone amplifiers consume some power even when the surroundings are silent.

3.1 Zero-Crossing VAD Algorithm

In our previous work, we proposed a low-power VAD hardware implementation using a single microphone[7]. This custom hardware uses a zero-crossing algorithm for the VAD. **Figure 3** portrays the zero-crossing algorithm, which is implemented on an FPGA in the ubiquitous sensor nodes as well.

The zero crossing is the first intersection between an input signal and an offset line after the signal crosses the trigger line: the high trigger line or 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 and outputs the beginning point and the end point of a speech segment. For the zero-crossing VAD to detect speech, the only requirement is catching the crossing over the trigger line and the offset line. A detailed speech signal is unnecessary. Conse-

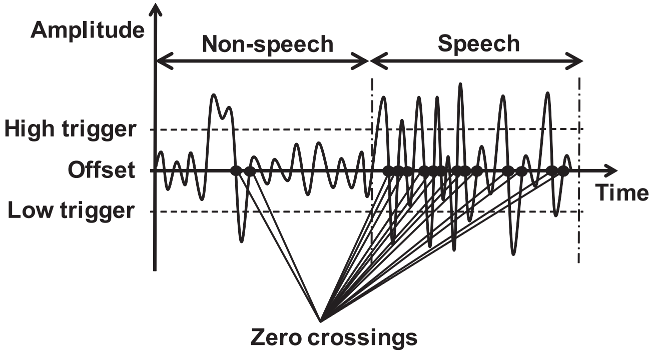


Fig. 3 Zero-crossing point example. The offset line shows the direct current (DC) component.

© 2011 Information Processing Society of Japan

quently, the sampling frequency and the number of bits can be reduced. Once the VAD module detects a speech signal, the main signal processor begins to run and the sampling frequency and the number of bits are increased to sufficient values. These parameters, which determine the analog digital converter (ADC) specifications, can be altered depending on the specific applications that are integrated with the system. As described herein, we adopt standard parameters: quality of 16 kHz sampling frequency and 16 bits per sample of most speech-recognition systems require continuous sensing [12]. Furthermore, only for the VAD algorithm, the sampling frequency is set to 2 kHz. The number of bits per sample is set to 10 bits. These values are sufficient to detect human speech, in which case only 3.49 μW is dissipated on a 0.18-μm CMOS process. By separating the low-power VAD module from the sound processing unit, the power manager can turn off the sound processing unit. A single microphone is sufficient to detect a signal. The remaining 15 microphones are turned off as well. Furthermore, not all VAD modules in all nodes need operate; VAD modules are activated in only a few nodes in the system.

3.2 Experimental Results

The SNR easily affects the zero-crossing VAD algorithms because the algorithms are based solely on changes in amplitude. For SNR dependencies of the VAD performance, we conduct experiments using various SNR environments of $-20 - 20$ dB. An input signal in this experiment is generated by adding recorded environmental sound, which is used for continual noise, to original speech data with gain control. In every SNR condition, we use identical 15-min speech data comprising 24 ATR phoneme balanced sentences [13]. The VAD algorithm frame length is 256 samples. In each SNR condition experiment, the number of VAD results is 7,030 samples. For this experiment, we counted the surplus and deficit VAD results. Each condition is defined as follows.

- ***False acceptance*** (***FA***): A case in which the VAD output is speech, although the input frame is non-speech.
- ***False rejection*** (***FR***): A case in which the VAD output is non-speech, although the input frame is speech.

Figures 4 and **5** respectively present results of ***FA*** VAD and ***FR*** VAD output quantities. The figures show that the power saving factor and stability of the zero-crossing VAD decrease according to SNR deterioration.

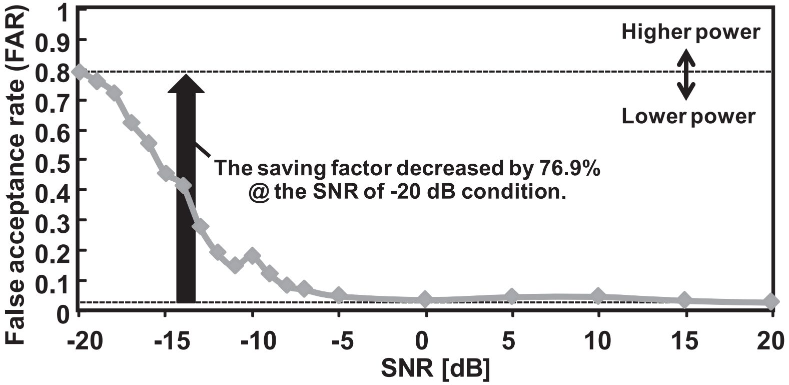


Fig. 4 The false acceptance rate (FAR) in VAD outputs using the number of non-speech frames of the recorded condition as normalized criteria.

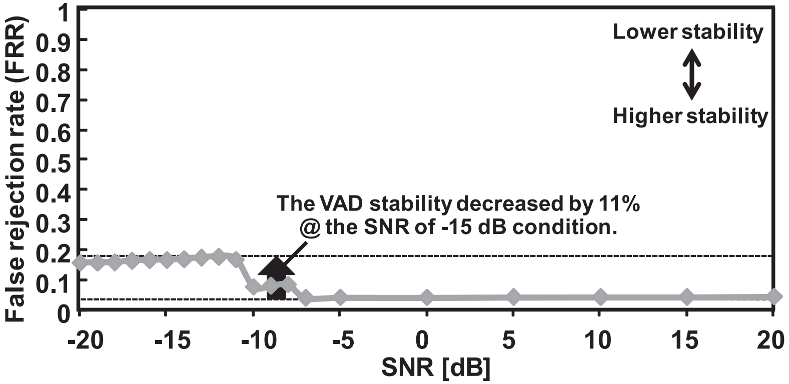


Fig. 5 The false rejection rate (FRR) in VAD outputs using the speech frames of the recorded condition as normalized criteria.

4. Proposed Sound Acquisition Scheme

As described in this paper, we examine microphone array networks specifically to obtain high-SNR sound data. To enhance the sound data, it is necessary to

© 2011 Information Processing Society of Japan

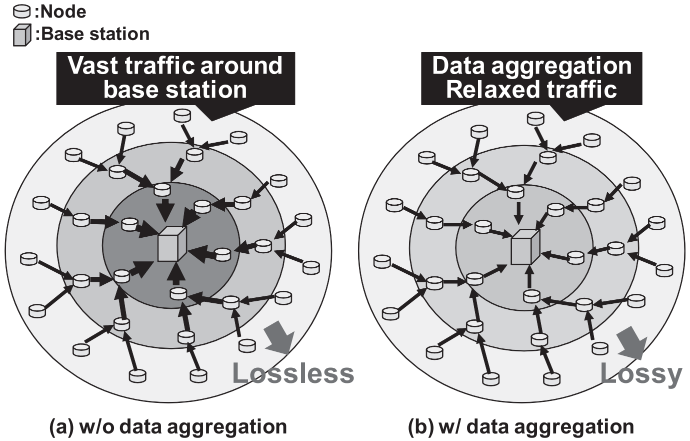


Fig. 6 Network traffic with (a) lossless and (b) lossy multi-hop networks.

gather sound data among numerous nodes and the network traffic is a bottleneck of a large system. Then we produce it as a multi-hop network. We propose a perfect aggregation solution that is specialized for obtaining high-SNR sound data in this section.

Some data aggregation techniques have been proposed to reduce network traffic for sensor networking. **Figure 6** presents network traffic with and without data aggregation. Without data aggregation, the network traffic is concentrated around the base station. An aggregation scheme must be chosen carefully according to the application. Data aggregation can be characterized as lossy or lossless[14]. Our aggregation method is chosen according to the former application. For applications such as reproduction of sound fields, lossless aggregation is suitable, but irreversible aggregation is sufficient for applications such as ours, which are intended solely to improve the sound SNR. Perfect aggregation[15] and beamforming[16] are lossy aggregations. With perfect aggregation, a sensor node aggregates the received data into one unit of data and then sends it to the next hop[17]. Therefore, perfect aggregation can reduce traffic on a grand scale.

4.1 Proposed Data Aggregation Scheme

In this subsection, we introduce the proposed perfect aggregation method. **Figure 7** presents an example of aggregation. In the figure, speech data acquired in nodes 1 and 2 are aggregated to single enhanced speech data in the aggregation node. Then the speech data are sent to the next node.

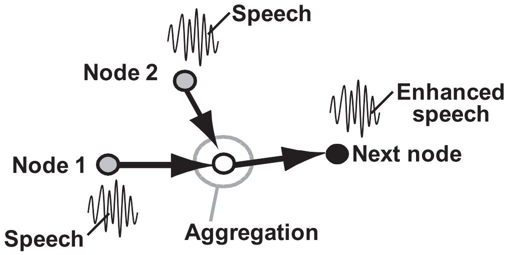


Fig. 7 Example of perfect aggregation among neighboring nodes.

To obtain high-SNR speech data, the aggregation algorithm must be eligible for a chosen sound-source enhancing method that lowers the noise signal level. Two types of major sound-source enhancing methods are geometric techniques, which use position information, and statistical techniques, which use no position information. For the proposed system, delay-and-sum beamforming, which is categorized as a geometric method, is chosen because the node positions in the network are known. This method produces less distortion than statistical techniques do. Fortunately, it requires only a small amount of computation. For distributed processing in sound source enhancement, it is easily applicable because it is based on summations (**Fig. 8**). The key point for delay-and-sum beamforming among distributed nodes is how to obtain time differences (W_i: phase differences in sound waves) among neighboring nodes.

Time differences among neighboring nodes are calculable from header information in a packet, which comprises a sound-source coordinate and a coordinate of each node. As a matter of course, the coordinate origin must be calibrated to a unique point. In aggregation using the timing data described above, all temporal speech data are adjusted by adding time differences and summing them to a single speech datum for uploading the signal. Consequently, high-SNR speech data can be acquired at the base station.

4.2 Three Dimensional Sound Source Localization

However, without a precise sound source coordinate, the delay-and-sum beamforming method does not operate effectively. For this reason, a basic sound-source

© 2011 Information Processing Society of Japan

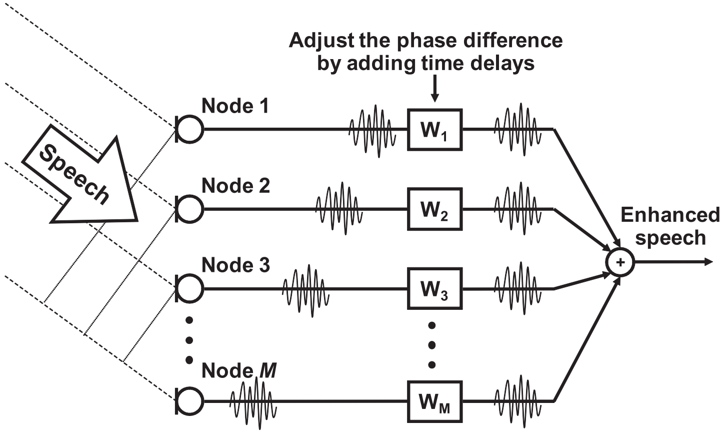


Fig. 8 Delay-and-sum beam-forming mechanism.

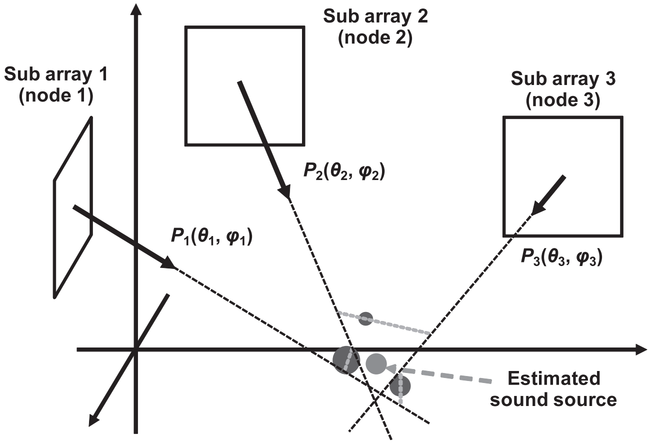


Fig. 9 Three-dimensional sound source localization.

localization algorithm with a high degree of accuracy is important to produce a perfect aggregation scheme. To achieve highly accurate sound source localization, we have proposed a hierarchical sound-source localization method [8)] based on the multiple signal classification (MUSIC) algorithm [18)–20)].

We adopt this MUSIC algorithm as the perfect aggregation method. We divide the localization into two layers—1) relative direction estimation within a node and 2) absolute location estimation—by exchanging results through the network. The MUSIC algorithm is chosen for direction estimation within a node because the number of microphones and their buffer memory on a node is limited; the MUSIC algorithm can achieve higher resolution using fewer microphones. To find a relative direction, the sound source probability $P(\theta, \phi)$ must be calculated for each node. Once the relative direction to the sound source is obtained, its information is transferred to neighboring nodes to proceed to the next step.

For the system, gathering and processing localization data are important to improve the localization accuracy. Distributed localization data obtained with the MUSIC algorithm [18)–20)] can be processed using a communication network in our system. We will localize the absolute sound source location in the network layer. The authors have proposed a calibration method with a three-dimensional coordinate of the sound source, as presented briefly in **Fig. 9** [8)]. First, the maximum $P(\theta, \phi)$ and corresponding θ and ϕ are calculated on each node using the MUSIC algorithm. We alternatively adopt the shortest line segment connecting two lines because we can usually find no exact intersection in the three-dimensional space. We presume a point that divides the shortest line segment by the ratios of $P(\theta, \phi)$ as an intersection. The sound source is localized by calculating the center of gravity as well, with the obtained intersections.

4.3 Simulation Results

We verified the hierarchical localization by simulation, assuming that an estimation result has a variation on every node. **Figure 10** presents an example of the experiments, for which the observed range is 12 m × 12 m. The localization accuracy is portrayed in **Fig. 11**. The localization error is smaller when the number of arrays is large and the direction estimation is precise. Results show that the effective means to make the localization accurate is to minimize the direction error. However, the number of sub-arrays does not give much impact, although the sub-array number strongly affects sound enhancement, as described later.

Although the coordinate data can be calibrated with nodes, the time stamp of each speech cannot be calibrated in this scheme. Time synchronization is an important issue for the delay-and-sum beamforming method. Timers of each sensor node, even among neighboring nodes, have dispersion because of vari-

© 2011 Information Processing Society of Japan

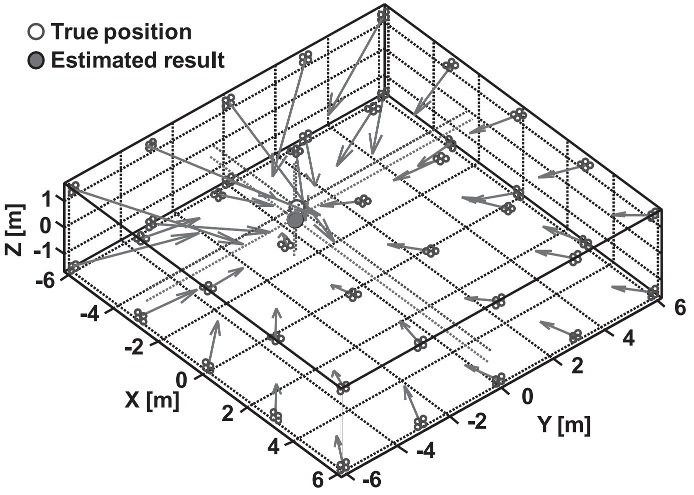


Fig. 10 Sound source localization experiment.

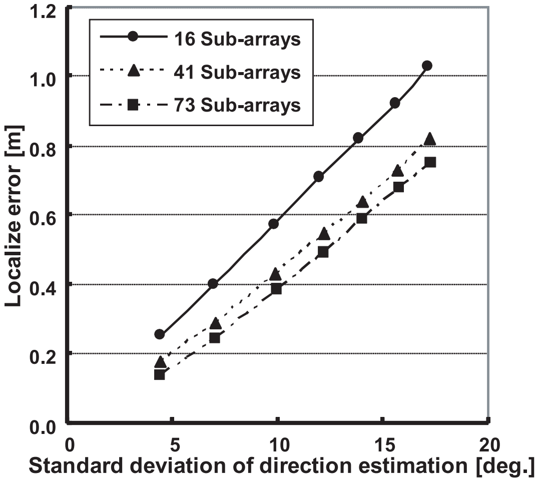


Fig. 11 Sound source localization accuracy.

ous environmental and device-origin effects. Therefore, the time synchronization method among nodes in sensor networks is important for the perfect aggregation scheme. Various means of time synchronization for a sensor network have been examined: reference Broadcast Synchronization (RBS) [21], Timing-sync Protocol for Sensor Networks (TPSN) [22], and Flooding Time Synchronization Protocol (FTSP) [23]. Using a time-synchronization protocol, infection to the SNR by the timer variation can be disregarded. For low-power multi-hop sensor networks such as microphone array networks, FTSP is the most suitable in terms of power consumption.

5. Implementation of the Microphone Array System

We implement the proposed perfect aggregation scheme to an actual sensor network with microphone arrays to verify the SNR performance. Regarding sound source enhancement, we use basic delay-and-sum beamforming (described in Section 4) both within a node and among nodes [11]. Therefore, the time accuracy between nodes gives a great impact on the final SNR of the sound source collected with the network.

For the actual design, we implemented an intelligent ubiquitous sensor node on a field-programmable gate array board (FPGA, SZ410, Suzaku; Atmark Techno Inc.) and microphones (ECM-C10; Sony Corp.). **Figure 12** portrays photographs of the prototype system. Each node performs sound source localization with its 16 microphones. Consequently, each node aggregates 16 sounds to a

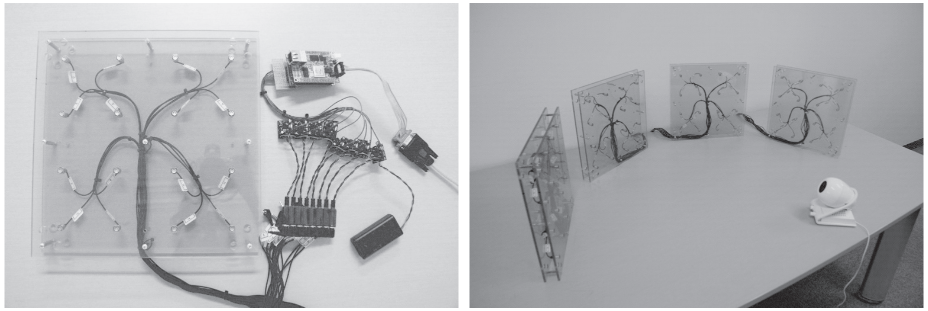


Fig. 12 System photographs: intelligent ubiquitous sensor node and microphone array comprising sub-arrays.

© 2011 Information Processing Society of Japan

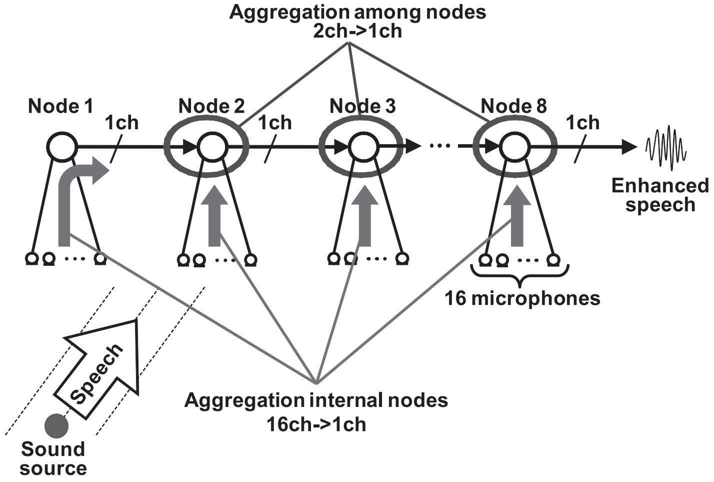


Fig. 13 Experiment diagram.

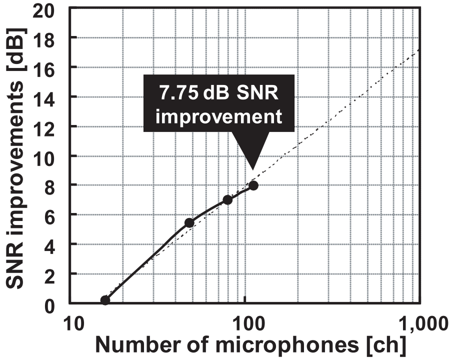


Fig. 14 SNR improvements vs. the number of microphones.

single sound using delay-and-sum beamforming, thereby enhancing the objective sound. Then the sound is transmitted to neighboring nodes. In this experiment, seven nodes are connected linearly, as shown in **Fig. 13**. They aggregate the data of one side to the other side. One aggregated audio datum, which has higher SNR, is obtained at the last node. All sounds from all 112 microphones are aggregated to one channel.

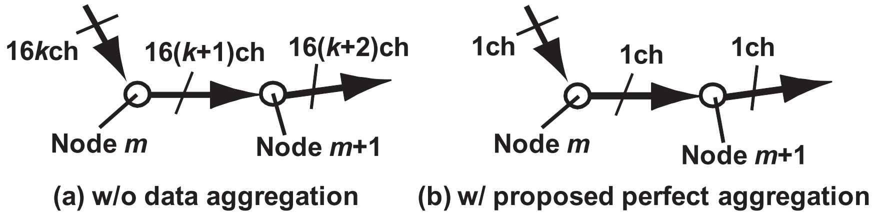


Fig. 15 Examples of traffic data sizes: (a) without and (b) with the proposed perfect data aggregations.

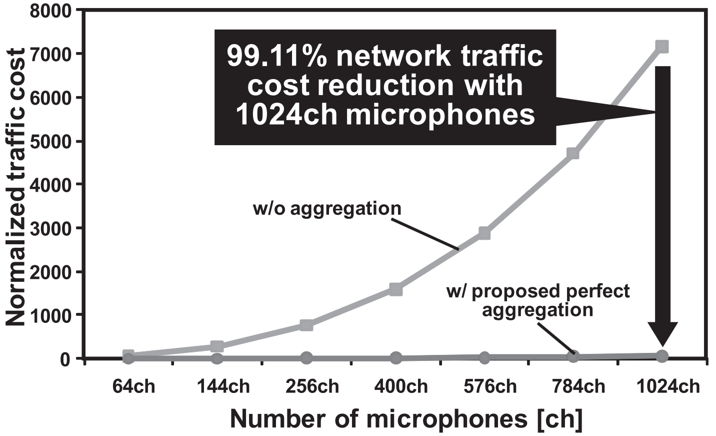


Fig. 16 Normalized traffic cost vs. the number of microphones.

Figure 14 shows that the SNR improvement of 7.75 dB was gained with 112 microphones. We expect to achieve 15 dB or greater improvement using several tens of sub-arrays and hundreds of microphones.

Next, we compared network-traffic costs with the proposed perfect aggregation and without data aggregation. **Figure 15** presents examples of the traffic data sizes with and without proposed perfect data aggregations. The network traffic is increased by 16 channels on every node without the data aggregation. This enables lossless sound acquisition and realizes applications such as the reproduction of sound fields, but results in heavy traffic (Fig. 15 (a)). Using the proposed perfect aggregation, the network traffic is always 1 channel (Fig. 15 (b)). This

© 2011 Information Processing Society of Japan

small-channel network produces lossy sound source acquisition. However, the sound-source localization algorithm and the sound-source enhancing algorithm achieve high SNR sound acquisition for an intended sound source.

Figure 16 shows normalized (the normalization criterion is the network cost in the proposed 32-ch perfect data aggregation) network costs with and without the proposed perfect data aggregations. For 1,024-ch microphones, the proposed perfect aggregation achieves 99.11% network traffic reduction, which demonstrates that the proposed scheme maintains the network traffic cost low consistently. It is applicable to a future larger-scale microphone array for a sound acquisition system.

6. Future Works

Our proposed system is implemented under the condition that the number of sound sources is one. Multiple sound sources and their separation are to be performed as future works. Compared to the single sound enhancement, the multiple sound separation necessitates sound source tracking because enhanced performance depends on the time-series direction and the system must recognize which sound source is the targeted source. In our proposed aggregation scheme, when considered with the multiple sound sources and their separations, the network traffic is increased linearly according to the number of sound sources: the issues are multiple sound source localization and tracking, plus increased network traffic for our future system.

Although the performance of sound source enhancement is known to be improved as the number of microphones is increased[6], power consumption and traffic costs are increased as well. Therefore, for actual implementation, the optimum deployment and power control (which nodes should be operated / which node can sleep) among nodes must be investigated further using a large-scale setup.

7. Conclusion

As described in this paper, we propose a perfect aggregation scheme that is specialized for sound acquisition systems comprising numerous microphones. The microphone array network using 16-microphone sub-arrays performed the following three operations in a node and a network: 1) low-power VAD to activate the entire node, 2) sound-source localization to find sound sources, and 3) sound-source enhancement to improve the SNR. We implemented an actual microphone array network that realizes a ubiquitous sound acquisition system, and verified that the proposed scheme reduces the network traffic and saves resources such as power and memory size.

Low-power VAD was implemented to manage the node's power consumption. The system uses very little power when speech is not active. The VAD module dissipates only 3.49 μW on a 0.18-μm CMOS process. Sound-source localization is processed with the distributed nodes. The proposed sound-source localization scheme uses a two-layered hierarchical algorithm. The experimental result of the sound-source enhancement shows SNR improvement of 7.75 dB using 112 microphones. The system achieves an SNR of 15 dB if the entire microphone network has more than several hundred microphones. We confirmed that the system achieves a 99.11% traffic amount reduction when using 1,024 microphones.

Acknowledgments The authors wish to thank Shinpei Soda of Kobe University for his ongoing effort for the project. This research was supported by the Semiconductor Technology Academic Research Center (STARC).

References

1) Akyildiz, I.F., Melodia, T. and Chowdhury, K.R.: A Survey on Wireless Multimedia Sensor Networks, *Journal of Computer Networks*, Vol.51, No.4, pp.921–960 (Mar. 2007).

2) Brandstein, M. and Ward, D.: *Microphone Arrays: Signal Processing Techniques and Applications*, Springer (2001).

3) Tamai, Y., Kagami, S., Mizoguchi, H., Sakaya, K., Nagashima, K. and Takano, T.: Circular Microphone Array for Meeting System, *Proc. IEEE Sensors*, Vol.2, pp.1100–1105 (Oct. 2003).

4) Wakabayashi, T., Takahashi, K. and Iwakura, H.: Independent Component Analysis using Large Microphone Array, IEICE Technical Report, Vol.102, No.322, pp.29–34 (Sep. 2002), in Japanese.

5) Silverman, H.F., Patterson III, W.R. and Flanagan, J.L.: The Huge Microphone Array, *Journal of IEEE Concurrency*, Vol.6, No.4, pp.36–46 (Oct.-Dec. 1998) and Vol.7, No.1, pp.32–47 (Jan.-Mar. 1999).

6) Weinstein, E., Steele, K., Agarwal, A. and Glass, J.: Loud: A 1020-Node Modular Microphone Array and Beamformer for Intelligent Computing Spaces, *MIT*,

© 2011 Information Processing Society of Japan

MIT/LCS Technical Memo, MIT-LCS-TM-642 (2004).

7) Noguchi, H., Takagi, T., Yoshimoto, M. and Kawaguchi, H.: An Ultra-Low-Power VAD Hardware Implementation for Intelligent Ubiquitous Sensor Networks, *Proc. IEEE Workshop on Signal Processing Systems* (*SiPS*), pp.214–219 (Oct. 2009).

8) Takagi, T., Noguchi, H., Kugata, K., Yoshimoto, M. and Kawaguchi, H.: Microphone Array Network for Ubiquitous Sound Acquisition, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp.1474–1477 (Mar. 2010).

9) Kugata, K., Takagi, T., Noguchi, H., Yoshimoto, M. and Kawaguchi, H.: Live Demonstration: Intelligent Ubiquitous Sensor Network for Sound Acquisition, *Proc. IEEE International Symposium on Circuits and Systems* (*ISCAS*), pp.1413–1417 (May 2010).

10) Noguchi, H., Takagi, T., Kugata, K., Yoshimoto, M. and Kawaguchi, H.: Low-Traffic and Low-Power Data-Intensive Sound Acquisition with Perfect Aggregation Specialized for Microphone Array Networks, *Proc. International Conference on Sensor Technologies and Applications* (*SENSORCOMM*), pp.157–162 (July 2010).

11) Benesty, J., Sondhi, M.M. and Huang, Y.: *Handbook of Speech Processing*, Springer (2007).

12) Fujinaga, T., Miura, K., Noguchi, H., Kawaguchi, H. and Yoshimoto, M.: Parallelized Viterbi Processor for 5,000-Word Large-Vocabulary Real-Time Continuous Speech Recognition FPGA System, *Proc. ISCA Annual Conference of International Speech Communication Association* (*Interspeech*), pp.1483–1486 (Sep. 2009).

13) Kobayashi, T., Itahashi, S., Hayamizu, S. and Takezawa, T.: ASJ Continuous Speech Corpus for Research, *Journal of Acoustical Society of Japan*, Vol.48, No.12, pp.888–893 (1992), in Japanese.

14) Abdelzaher, T.F., He, T. and Stankovic, J.A.: Feedback Control of Data Aggregation in Sensor Networks, *Proc. IEEE Conference on Decision and Control* (*CDC*), Vol.2, pp.1490–1495 (Dec. 2004).

15) Intanagonwiwat, C., Estrin, D., Govindan, R. and Heidemann, J.: Impact of Density on Data Aggregation in Wireless Sensor Networks, *Proc. 22nd International Conference on Distributed Computing Systems*, pp.457–458 (Nov. 2001).

16) Wang, A., Heinzelman, W.B., Sinha, A. and Chandrakasan, A.P.: Energy-Scalable Protocols for Battery-Operated MicroSensor Networks, *Journal of VLSI Signal Processing systems*, Vol.29, No.3, pp.223–237 (Nov. 2001).

17) Zhao, J., Govindan, R. and Estrin, D.: Computing Aggregates for Monitoring Wireless Sensor Networks, *Proc. IEEE International Workshop on Sensor Network Protocols and Applications*, pp.139–148 (May 2003).

18) Asano, F., Asoh, H. and Matsui, T.: Sound Source Localization and Signal Separation for Office Robot (Jijo-2), *Proc. IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems* (*MFI 1999*), pp.243–248 (Aug. 1999).

19) Tanaka, H. and Kobayashi, T.: Estimating Positions of Multiple Adjacent Speakers Based on MUSIC Spectra Correlation using a Microphone Array, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Vol.5, pp.3045–3048 (May 2001).

20) Nakadai, K., Nakajima, H., Murase, M., Kaijiri, S., Yamada, K., Nakamura, T., Hasegawa, Y., Okuno, H.G. and Tsujino, H.: Robust Tracking of Multiple Sound Sources by Spatial Integration of Room and Robot Microphone Arrays, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Vol.4, pp.929–932 (May 2006).

21) Elson, J., Girod, L. and Estrin, D.: Fine-Grained Network Time Synchronization using Reference Broadcasts, *Proc. 5th ACM SIGOPS Symposium on Operating Systems Design and Implementation* (*OSDI'02*), pp.147–163 (Dec. 2002).

22) Ganeriwal, S., Kumar, R. and Srivastava, M.B.: Timing-Sync Protocol for Sensor Networks, *Proc. 1st ACM Conference on Embedded Networked Sensor Systems* (*SenSys'03*), pp.138–149 (Nov. 2003).

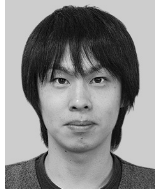
23) Maroti, M., Kusy, B., Simon, G. and Ledeczi, A.: The Flooding Time Synchronization Protocol, *Proc. 2nd ACM Conference on Embedded Networked Sensor Systems* (*SenSys'04*), pp.39–49 (Nov. 2004).

(Received May 22, 2010)
(Accepted November 5, 2010)
(Released March 9, 2011)

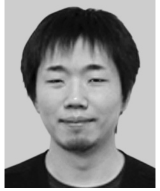
Hiroki Noguchi received his B.E. and M.E. degrees in Computer and Systems Engineering in 2006 and 2008, respectively from Kobe University, Hyogo, Japan, where he is currently earning a Ph.D. degree. His research interests are low-power SRAM designs, multimedia/ubiquitous systems and digital signal processing architectures, which include speech-recognition for handheld, image-recognition for wearable computing, and mixed integer programming for real-time robotics controlling, and their low-power hardware implementation. He is a student member of IEICE and IEEE.

© 2011 Information Processing Society of Japan

Tomoya Takagi earned a B.E. degree in Computer and Systems Engineering from Kobe University, Hyogo, Japan, in 2009. He is a master course student at Kobe University. His current research is about microphone array network and its low-power implementation. He is a student member of IEEE.



Koji Kugata earned a B.E. degree in Computer and Systems Engineering from Kobe University, Hyogo, Japan, in 2010. Currently, he is a master course student at Kobe University. His current research is related to low-voltage operation circuits for digital signal processing.



Shintaro Izumi received his B.E. and M.E. degrees in Computer Science and Systems Engineering in 2007 and 2008, respectively from Kobe University, Hyogo, Japan. Currently, he is a Ph.D. course student and a JSPS research fellow at Kobe University. His current research interests include communication protocols, low-power VLSI design, and wireless sensor networks. He is a student member of IEEE.

Masahiko Yoshimoto earned his B.S. degree in Electronic Engineering from Nagoya Institute of Technology, Nagoya, Japan, in 1975, and an M.S. degree in Electronic Engineering from Nagoya University, Nagoya, Japan, in 1977. He earned a Ph.D. degree in Electrical Engineering from Nagoya University, Nagoya, Japan in 1998. He joined the LSI Laboratory, Mitsubishi Electric Corp., Itami, Japan, in April 1977. During 1978–1983 he was engaged in the design of NMOS and CMOS static RAM including a 64 K full CMOS RAM with the world's first divided-wordline structure. From 1984, he was involved in research and development of multimedia ULSI systems for digital broadcasting and digital communication systems based on MPEG2 and MPEG4 Codec LSI core technology. Since 2000, he has been a Professor of the Deptartment of Electrical and Electronic Systems Engineering at Kanazawa University, Japan. Since 2004, he has been a Professor of the Deptartment of Computer and Systems Engineering at Kobe University, Japan. His current activities are focused on research and development of multimedia and ubiquitous media VLSI systems including an ultra-low-power image compression processor and a low-power wireless interface circuit. He holds 70 registered patents. He served on the Program Committee of IEEE International Solid State Circuit Conference from 1991 to 1993. Additionally, he served as a Guest Editor for special issues on Low-Power System LSI, IP, and Related Technologies of IEICE Transactions in 2004. He received R&D100 awards in 1990 and 1996 from R&D Magazine for development of the DISP and development of a real-time MPEG2 video encoder chipset, respectively.

© 2011 Information Processing Society of Japan

Hiroshi Kawaguchi received his B.E. and M.E. degrees in Electronic Engineering from Chiba University, Chiba, Japan, in 1991 and 1993, respectively, and earned a Ph.D. degree in Engineering from the University of Tokyo, Tokyo, Japan, in 2006. He joined Konami Corporation, Kobe, Japan, in 1993, where he developed arcade entertainment systems. He moved to the Institute of Industrial Science, the University of Tokyo, as a Technical Associate in 1996, and was appointed as a Research Associate in 2003. In 2005, he moved to Kobe University, Kobe, Japan. Since 2007, he has been an Associate Professor with the Department of Information Science at that university. He is also a Collaborative Researcher with the Institute of Industrial Science, the University of Tokyo. His current research interests include low-voltage SRAM, RF circuits, and ubiquitous sensor networks. Dr. Kawaguchi was a recipient of IEEE ISSCC 2004 Takuo Sugano Outstanding Paper Award and IEEE Kansai Section 2006 Gold Award. He has served as a Design and Implementation of Signal Processing Systems (DISPS) Technical Committee Member for IEEE Signal Processing Society, as a Program Committee Member for IEEE Custom Integrated Circuits Conference (CICC) and IEEE Symposium on Low-Power and High-Speed Chips (COOL Chips), and as a Guest Associate Editor of IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences and IPSJ Transactions on System LSI Design Methodology (TSLDM). He is a member of IEEE, ACM, IEICE, and IPSJ.

© 2011 Information Processing Society of Japan