# Low-Traffic and Low-Power Data-Intensive Sound Acquisition with Perfect Aggregation Specialized for Microphone Array Networks

Hiroki Noguchi, Tomoya Takagi, Koji Kugata, Masahiko Yoshimoto, and Hiroshi Kawaguchi
Department of Computer and Systems Engineering, Kobe University
1-1, Rokkodai, Nada, Kobe, 657-8501 Japan
e-mail: {h-nog, takagi, kugata, yosimoto, kawapy}@cs28.cs.kobe-u.ac.jp

***Abstract*—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 separation. The three operations are distributed among nodes. Using the distributed network, we achieve a low-traffic data-intensive array network. To manage nodes' power consumption, VAD is implemented. Consequently, the system uses little power when speech is not active. For sound localization, a network-connected multiple signal classification (MUSIC) algorithm is used. The sound separation system can improve a signal-noise ratio (SNR) by 7.75 dB using 112 microphones. Network traffic is reduced by 99.11% when using 1024 microphones.**

***Keywords—microphone array; ubiquitous sensing; sensor network; low-power system; perfect aggregation.***

## I. INTRODUCTION

In recent years, improvement in information processing technology has produced real-time sound processing systems using microphone arrays. A microphone array can localize sound sources and separate multiple sources using spatial information of the acquired sounds. The computational effort of these operations increases polynomially with the number of microphones, but the operating performance is known to increase as well [1]. To reduce the increased power of a microphone array and to satisfy recent demands for ubiquitous sound acquisition, it is necessary to realize a large sound processing system at low power.

Huge microphone arrays have been widely investigated: arrays have been built at Tokyo University of Science (128 ch) [2], the University of Electro-Communication (156 ch) [3], Brown University and Rutgers University (512 ch) [4], [5], and the Massachusetts Institute of Technology (1,020 ch) [1]. However, the problems of increasing computation, power consumption, and cost make their practical use difficult, particularly in terms of sound-data acquisition. The main problem of conventional microphone array systems is that all the microphones are connected to a single base station (sound server). Reducing the amount of transmission should be accomplished by introducing multi-hop networking.

To implement a microphone array as a real ubiquitous sound acquisition system, we have proposed division of the huge array into sub-arrays to produce a multi-hop network: an intelligent ubiquitous sensor network (IUSN) [6–8]. The sub-array nodes can be set up on the walls and ceiling of a room. Their performance can be improved easily merely by increasing the node number, but communication among nodes does not increase much in our system.

If more than 1,000 microphones are used to collect the data, then the signal-noise ratio (SNR) can be improved remarkably, but the network traffic would burst out. Therefore, each relay node on a routing path must store all temporal multi-channel sound data that the node receives, but not send it. This engenders a large-size buffer memory and large total power dissipation in the system. For that reason, some breakthrough network solution is needed to reduce the network traffic, even in a large-scaled network.

Some data aggregation techniques have been proposed to reduce network traffic for sensor networking. Fig. 1 presents network traffic with and without data aggregation. Without data aggregation, the network traffic is concentrated around the base station. An aggregation scheme should be chosen carefully according to the application. Data aggregation is classifiable as lossy and lossless [9]. Our aggregation method is chosen according to the former application. For applications such as reproduction of sound fields, lossless aggregation is suitable. However, irreversible aggregation is sufficient for applications such as ours, which are solely intended to improve the SNR of sound. Perfect aggregation [10] and beam forming [11] 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 [12]. Therefore, perfect aggregation can reduce traffic on a grand scale.

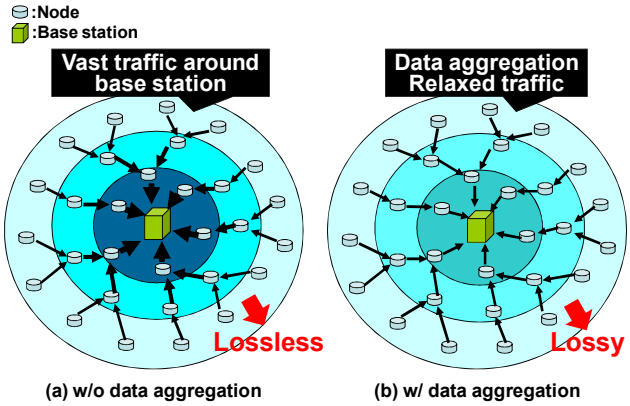


Figure 1. Network traffic with (a) lossless and (b) lossy multi-hop networks.

978-0-7695-4096-2/10 $26.00 © 2010 IEEE
DOI 10.1109/SENSORCOMM.2010.32

As described in this paper, we specifically examine microphone array networks to obtain high-SNR sound data. Then we produce it as a multi-hop network. We propose a perfect aggregation solution that is specialized for obtaining high-SNR sound data. We then demonstrate that the network traffic is reduced dramatically. Consequently, our proposed microphone array system is scalable.

## II. Proposed Data Aggregation Scheme

In this section, we introduce the proposed perfect aggregation method. Fig. 2 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.

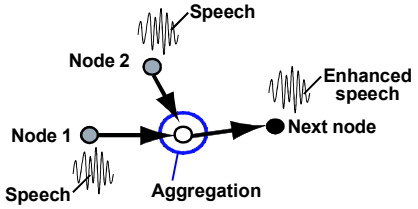


Figure 2. Example of perfect aggregation among neighboring nodes.

To obtain high-SNR speech data, the aggregation algorithm must be eligible for a chosen sound-source separation method that lowers the noise signal level. Two types of major sound-source separation methods are geometric techniques, which use position information, and statistical techniques, which use no position information. For the proposed system, delay-and-sum beam forming, 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 separation, it can be applied easily because it is based on summations (Fig. 3). The key point for delay-and-sum beam forming among distributed nodes is how to obtain time differences ($W_i$: phase differences in sound waves) among neighboring nodes.

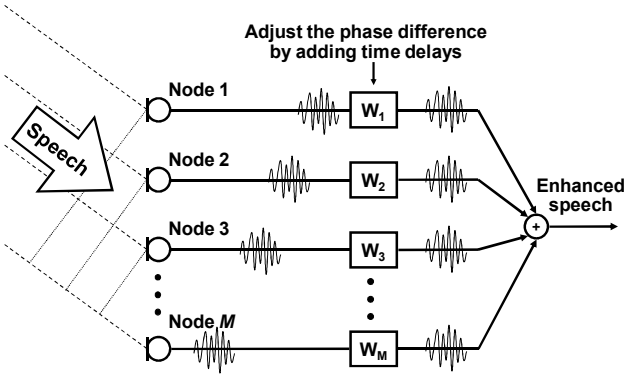


Figure 3. Delay-and-sum beam-forming mechanism.

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.

However, without a precise sound source coordinate, the delay-and-sum beam-forming method does not operate effectively. For this reason, a basic sound-source 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 already proposed a hierarchical sound-source localization method [7] based on the multiple signal classification (MUSIC) algorithm [13–15].

As described herein, we will divide the localization into two layers: 1) relative direction estimation within a node, and 2) absolute location estimation by exchanging the results through the network. The MUSIC algorithm for the relative direction estimation is based on subspace techniques for estimating the directions of arrival (DOAs) of multiple signal sources. We assume an array comprising $N$ microphones that receive signals from $L$ ($L < N$) sources. The $N \times 1$ array output at time $t$ can be modeled as

$$\mathbf{x}(t) = \mathbf{A}(\boldsymbol{\theta})\mathbf{s}(t) + \mathbf{n}(t), \tag{1}$$

where $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_L]^T$ is the DOA vector, $\boldsymbol{s}(t)$ is the $L \times 1$ vector of signal waveforms, $\boldsymbol{n}(t)$ is the $N \times 1$ vector of noise and interference, and

$$\mathbf{A}(\boldsymbol{\theta}) = [\mathbf{a}(\theta_1), \cdots, \mathbf{a}(\theta_L)], \tag{2}$$

is the $N \times L$ signal steering matrix [14, 15]. We presume that no coherent signals exist and that the noise is spatially white. Consequently, the $N \times N$ array covariance matrix can be written as

$$\mathbf{R}_{\mathbf{x}} = \mathrm{E}\{\mathbf{x}(t)\mathbf{x}^H(t)\} = \mathbf{A}\mathbf{R}_{\mathbf{s}}\mathbf{A}^H + \sigma^2\mathbf{I}, \tag{3}$$

where $\mathbf{R}_{\mathbf{s}} = \mathrm{E}\{\mathbf{s}(t)\mathbf{s}^H(t)\}$ is the source covariance matrix, $\sigma^2$ is the sensor noise variance, and $\boldsymbol{I}$ is the identity matrix [16, 17]. Using a unitary matrix $\boldsymbol{E}$, the $\boldsymbol{R_x}$ can be transformed orthogonally as

$$\mathbf{E}^H\mathbf{R}_{\mathbf{x}}\mathbf{E} = \mathrm{diag}[\lambda_1, \cdots, \lambda_L, 0, \cdots, 0] + \sigma^2\mathbf{I}, \tag{4}$$

where $\lambda_1, \cdots, \lambda_L$ are eigenvalues of $\boldsymbol{R_x}$. The unitary matrix $\boldsymbol{E}$ can be represented as $\mathbf{E} = [\mathbf{e}_1, \cdots, \mathbf{e}_N]$, where $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_N$ are eigenvectors of $\boldsymbol{R_x}$. Thereby, eq. (4) can be expressed as

$$\mathbf{R}_{\mathbf{x}} = \sum_{n=1}^{L}(\lambda_n + \sigma^2)\mathbf{e}_n\mathbf{e}_n^H + \sigma^2\sum_{n=L+1}^{N}\mathbf{e}_n\mathbf{e}_n^H, \tag{5}$$

using the relation $\mathbf{E}\mathbf{E}^H = \mathbf{E}^H\mathbf{E} = \mathbf{I}$. Between (3) and (5), $\mathbf{a}(\theta_1), \cdots, \mathbf{a}(\theta_L)$ and $\mathbf{e}_{L+1}, \cdots, \mathbf{e}_N$ are orthogonal. Consequently, the source location can be found by plotting the following quantity as a function of $\theta$:

$$P(\theta) = \frac{1}{\mathbf{a}(\theta)^H\sum_{n=L+1}^{N}\mathbf{e}_n\mathbf{e}_n^H\mathbf{a}(\theta)}. \tag{6}$$

Fig. 4 presents an example of the result: $P(\theta)$.

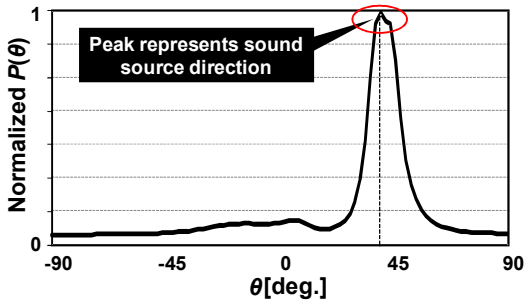


Figure 4. Example of the result from one-dimensional sound source localization.

We adopt the above MUSIC algorithm as the perfect aggregation method. 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 $P(\theta, \varphi)$ must be calculated on 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.

We will localize the absolute sound source location in the network layer. The authors have already proposed a calibration method with a three-dimensional coordinate of the sound source, as presented briefly in Fig. 5 [7]. First, the maximum $P(\theta, \varphi)$ and corresponding $\theta$ and $\varphi$ 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 $P(\theta, \varphi)$ as an intersection. The sound source is localized by calculating the center of gravity as well, with the obtained intersections.

We verified the hierarchical localization by simulation, assuming that an estimation result has a variation on every node. Fig. 6 presents an example of the experiments. The localization accuracy is portrayed in Fig. 7. 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. (The number of sub-arrays does strongly affect sound separation, 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 beam-forming method. Timers of each sensor node, even among neighboring nodes, have dispersion by various environmental and device-origin effects. For that reason, 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) [18], Timing-sync Protocol for Sensor Networks (TPSN) [19], and Flooding Time Synchronization Protocol (FTSP) [20]. 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.

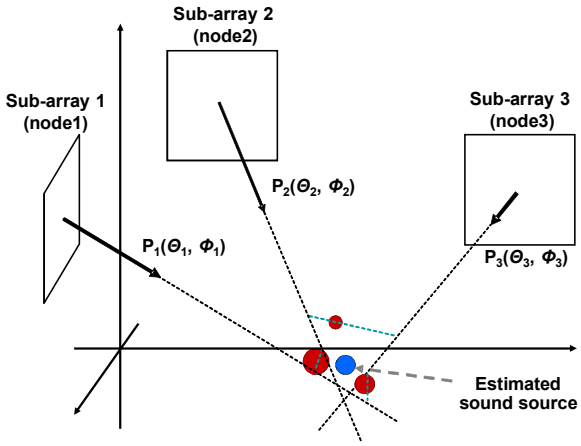


Figure 5. Three-dimensional sound source localization.

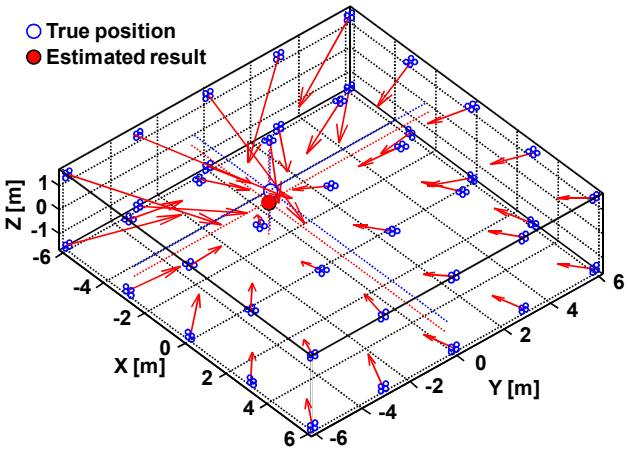


Figure 6. Sound source localization experiment.

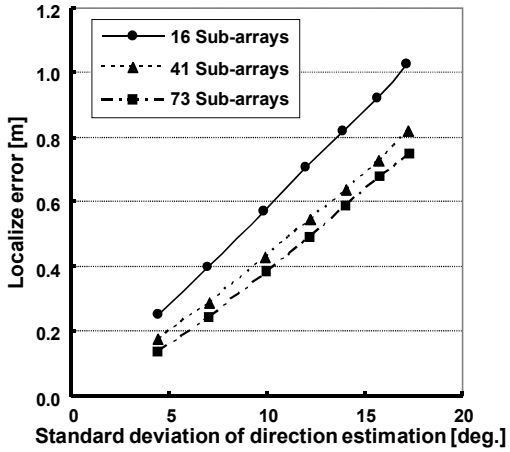


Figure 7. Sound source localization accuracy.

## III. Intelligent Ubiquitous Sensor Network and Its Node

This section describes implementation of the proposed perfect aggregation scheme to a microphone array system.

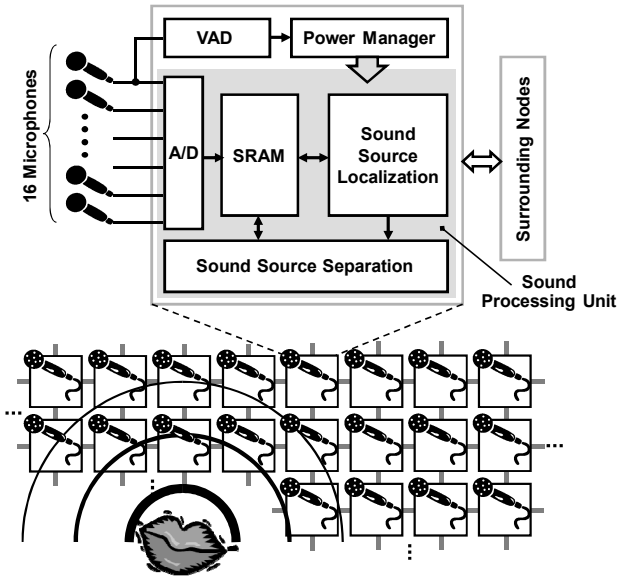


Figure 8. Intelligent ubiquitous sensor network (IUSN) and block diagram of a sub-array node.

Fig. 8 shows a brief description of the proposed IUSN and a functional block diagram of a sub-array node. Sixteen-microphone inputs are digitized with 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. The power management is fundamentally required because enormous microphones waste much power whereas 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 [6]. 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 tuned off as well. Furthermore, not all VAD modules in all nodes need operate. The VAD modules in a limited number of nodes are merely activated in the system.

Fig. 9 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 find sound sources, and 3) sound-source separation to enhance the sound. 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 achieves a huge microphone array through interaction with surrounding nodes. Therefore, the computations can be distributed among nodes. The system has scalability in terms of the number of microphones. Each node preprocesses acquired sound data; then only compressed data—localized and separated sound––are communicated.

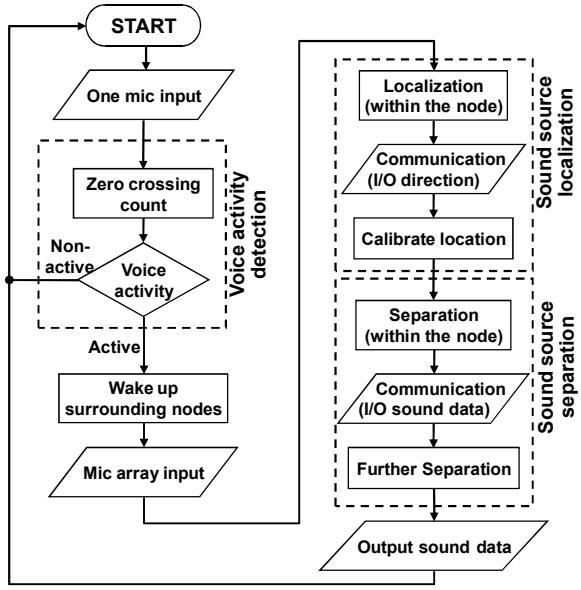


Figure 9. Flow chart of intelligent ubiquitous sensor nodes.

As a real design, we implemented the intelligent ubiquitous sensor node on a field-programmable gate array board (FPGA, SZ410, Suzaku; Atmark Techno Inc.) and microphones (ECM-C10; Sony Corp.). Fig. 10 portrays photographs of the prototype system.

The next section presents discussion of the performances and accuracies of the sound-source localization and sound-source separation 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 with the MUSIC algorithm can be processed using a communication network in our system. Regarding the sound-source separation, we use basic delay-and-sum beam forming both within a node and among nodes [21]. Therefore, the time accuracy between nodes strongly affects the final SNR of the sound source collected with the network.

## IV. Implementation of the Microphone Array System with the Proposed Aggregation Scheme

We implement the proposed perfect aggregation scheme to an actual sensor network with microphone arrays to verify the SNR performance. Each node operates the sound source localization with its 16 microphones. Consequently, each node aggregates 16 sounds to a single sound using delay-and-sum beam-forming enhancing the objective sound. Then the sound is transmitted to neighboring nodes. In this experimentation, seven nodes are connected linearly, as illustrated in Fig. 11. Then 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.

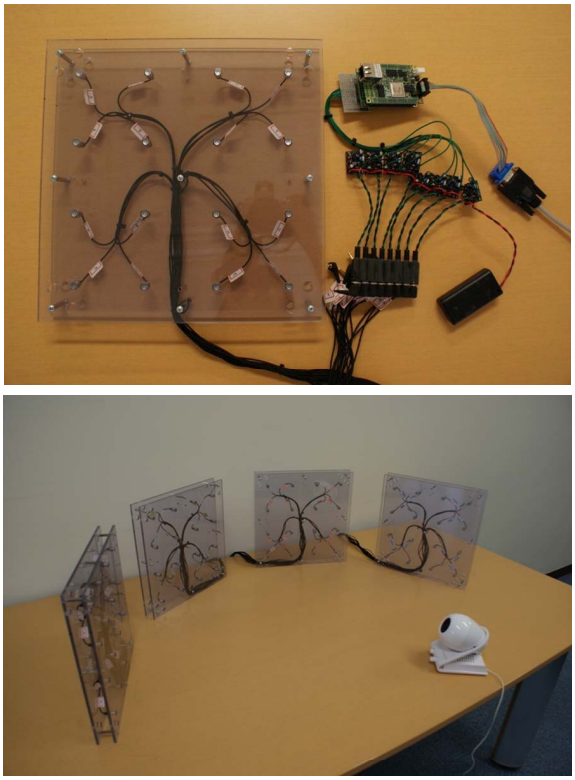

Figure 10. System photographs: intelligent ubiquitous sensor node and microphone array comprising sub-arrays.

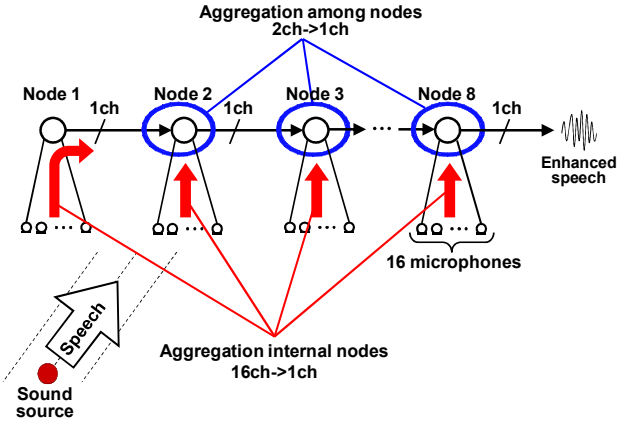


Figure 11. Experiment diagrams.

Fig. 12 shows that the SNR improvement of 7.75 dB was gained with 122 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. Fig. 13 shows examples of the traffic data sizes with and without proposed perfect data aggregations. The network traffic is increased by 16 channels on every node if 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. 13(a)). On the other hand, using the proposed perfect aggregation, the network traffic is always 1 channel (Fig. 13(b)). This small-channel network means a lossy sound source acquisition. However, the sound-source localization algorithm and the sound-source separation algorithm achieve high SNR sound acquisition for an intended sound source.

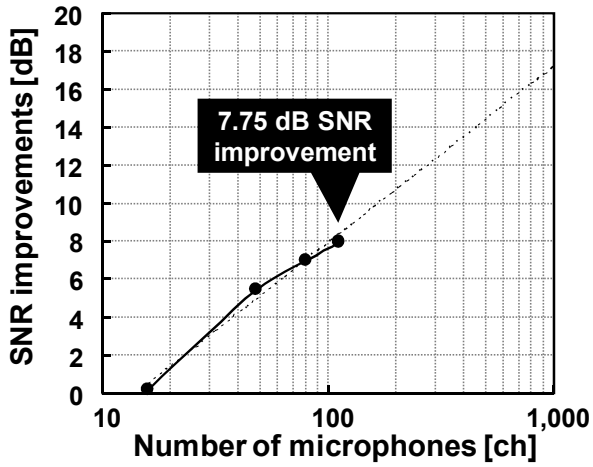


Figure 12. SNR improvements vs. the number of microphones.

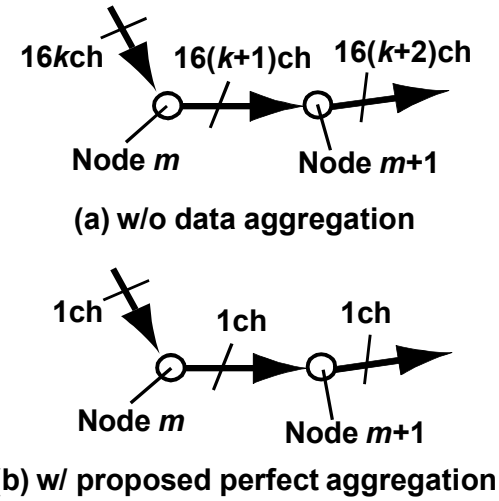


Figure 13. Examples of traffic data sizes: (a) without and (b) with the proposed perfect data aggregations.

Fig. 14 shows normalized (the criterion for normalization is the network cost in the proposed 32-ch perfect data aggregation) network costs with and without the proposed perfect data aggregations. For 1024-ch microphones, the proposed perfect aggregation achieves 99.11% network traffic reduction, which demonstrates that the proposed scheme keeps the network traffic cost low consistently. It is applicable to a future larger-scale microphone array for a sound acquisition system.

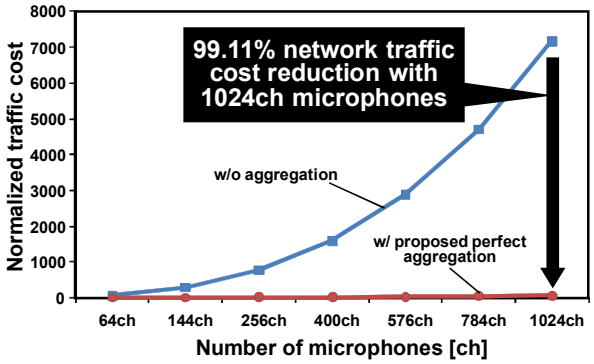


Figure 14. Normalized traffic cost vs. the number of microphones.

## V. CONCLUSION

As described in this paper, we propose a novel 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 separation to enhance the sound. We implemented an actual microphone array network that realizes the 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 achieves low 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 separation shows SNR improvement of 7.75 dB using 112 microphones. We confirmed that the system achieves a 99.11% traffic amount reduction when using 1024 microphones.

## ACKNOWLEDGMENT

This research has been supported by the Semiconductor Technology Academic Research Center (STARC).

## REFERENCES

[1] E. Weinstein, K. Steele, A. Agarwal, and J. Glass, "Loud: A 1020-node modular microphone array and beamformer for intelligent computing spaces," *MIT, MIT/LCS Technical Memo*, MIT-LCS-TM-642, 2004.

[2] M. Takamiya, T. Sekitani, Y. Miyamoto, Y. Noguchi, H. Kawaguchi, T. Someya, and T. Sakurai, "Design Solutions for a Multi-Object Wireless Power Transmission Sheet Based on Plastic Switches," in Proc. of *IEEE ISSCC 2007*, pp. 362-363, 2007.

[3] Y. Tamai, S. Kagami, H. Mizoguchi, K. Sakaya, K. Nagashima, and T. Takano, "Circular microphone array for meeting system," in Proc. of *IEEE Sensors*, vol. 2, pp. 1100-1105, Oct. 2003.

[4] M. Brandstein and D. Ward, "Microphone Arrays: Signal Processing Techniques and Applications," *Springer*, 2001.

[5] H. F. Silverman, W. R. Patterson III, and J. L. Flanagan, "The Huge Microphone Array," Oct. 1996.

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

[7] T. Takagi, H. Noguchi, K. Kugata, M. Yoshimoto, and H. Kawaguchi, "Microphone Array Network for Ubiquitous Sound Acquisition," in Proc. of *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1474-1477, Mar. 2010.

[8] K. Kugata, T. Takagi, H. Noguchi, M. Yoshimoto, and H. Kawaguchi, "Live Demonstration: Intelligent Ubiquitous Sensor Network for Sound Acquisition," in Proc. of *IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2010. (To be demonstrated)

[9] T.F. Abdelzaher, T. He, and J.A. Stankovic, "Feedback control of data aggregation in sensor networks," in Proc. of *IEEE Conference on Decision and Control*, Dec. 2004.

[10] C. Intanagonwiwat, D. Estrin, R. Govindan, and J. Heidemann, "Impact of density on data aggregation in wireless sensor networks," in Proc. of *22nd International Conference on Distributed Computing Systems*, Nov. 2001.

[11] A. Wang, W.B. Heinzelman, A. Sinha, and A.P. Chandrakasan, "Energy-scalable for battery-operated microsensor networks," *Journal of VLSI Signal Processing*, vol. 29, pp. 223–237, Nov. 2001.

[12] J. Zhao, R. Govindan, and D. Estrin, "Computing aggregates for monitoring wireless sensor networks," in Proc. of *IEEE International Workshop on Sensor Network Protocols and Applications*, May 2003.

[13] F. Asano, H. Asoh, and T. Matsui, "Sound source localization and signal separation for office robot (Jijo-2)," in Proc. of *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 1999)*, pp. 243-248, Aug. 1999.

[14] H. Tanaka and T. Kobayashi, "Estimating positions of multiple adjacent speakers based on MUSIC spectra correlation using a microphone array," in Proc. of *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2001.

[15] K. Nakadai, H. Nakajima, M. Murase, S. Kaijiri, K. Yamada, T. Nakamura, Y. Hasegawa, H.G. Okuno, and H. Tsujino, "Robust Tracking of Multiple Sound Sources by Spatial Integration of Room And Robot Microphone Arrays," in Proc. of *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2006.

[16] R.O. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Transactions of Antennas Propagation*, vol. 34, pp. 276-280, Mar. 1986.

[17] G. Bienvenu and L. Kopp, "Adaptivity to Background Noise Spatial Coherence for High resolution Passive Methods," in Proc. of *IEEE ICASSP'80*, pp. 307-310, Denver, USA, Apr. 1980.

[18] J. Elson, L. Girod, and D. Estrin, "Fine-grained network time synchronization using reference broadcasts," in Proc. of *5th Symposium on Operating Systems Design and Implementation (OSDI'02), Boston, Massachusetts*, pp. 147-163, 2002.

[19] S. Ganeriwal, R. Kumar, and M.B. Srivastava, "Timing-sync protocol for sensor networks," in Proc. of *1st ACM Conference on Embedded Networked Sensor Systems (SenSys'03), Los Angeles, California*, pp. 138-149, 2003.

[20] M. Maroti, B. Kusy, G. Simon, and A. Ledeczi, "The flooding time synchronization protocol," in Proc. of *2nd ACM Conference on Embedded Networked Sensor Systems (SenSys'04), Baltimore, Maryland*, pp. 39-49, 2004.

[21] J. Benesty, M.M. Sondhi, and Y. Huang, "Handbook of Speech Processing," *Springer*, 2007.