MICROPHONE ARRAY NETWORK FOR UBIQUITOUS SOUND ACQUISITION

Tomoya Takagi, Hiroki Noguchi, Koji Kugata, Masahiko Yoshimoto, and Hiroshi Kawaguchi

Graduate School of Engineering, Kobe University
1-1 Rokkodai, Nada, Kobe, Hyogo, 657-8501 Japan
takagi@cs28.cs.kobe-u.ac.jp

ABSTRACT

We propose a microphone array network that realizes ubiquitous sound acquisition. Nodes with 16 microphones are connected to form a huge sound acquisition system that carries out VAD, sound source localization and separation. The three operations are distributed among nodes. The VAD is implemented to manage power consumption. Consequently, the system consumes little power when speech is not active. The VAD module uses only 2.1 mW. The system can improve an SNR by 7.75 dB using 112 microphones.

Index Terms— Microphone array, ubiquitous sensing, sensor network, distribution network, low-power system

1. INTRODUCTION

In recent years, information processing technology improvements have realized real-time sound processing systems with microphone arrays. A microphone array can localize sound sources and separate multiple sources using the acquired sounds' spatial information. The computational effort of these operations increases polynomially with the number of microphones, but the performance of these operations is known to improve concomitantly [1]. To reduce the increasing power of a microphone array and to satisfy the recent demand for ubiquitous sound acquisition, it is necessary to realize a low-power, large sound-processing system.

Huge microphone arrays have been widely researched 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, obstacles to their practical use persist: increasing computation, power consumption, and cost.

To implement a microphone array as an actual ubiquitous sound acquisition system, we propose to divide the huge array into sub-arrays and produce a network: an intelligent ubiquitous sensor network (IUSN). The sub-array nodes can be set up on a room's walls and ceiling. The performance can be improved by increasing the nodes, but the communication between nodes does not increase so much in our system. As described herein, we present our IUSN solution to the problems listed above.

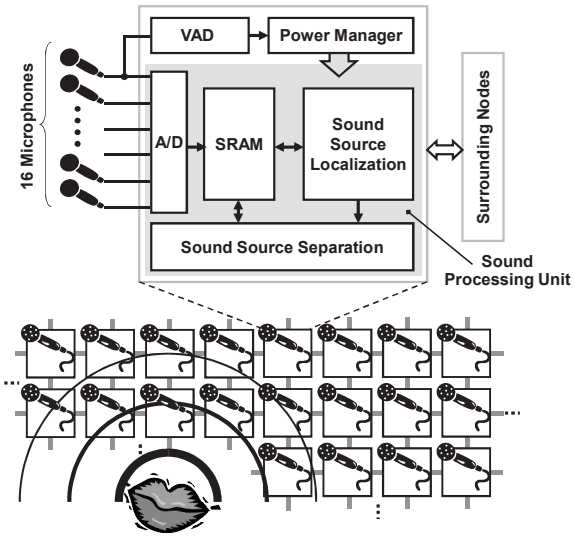

Fig. 1 Intelligent ubiquitous sensor network (IUSN) and block diagram of a sub-array node.

2. INTELLIGENT UBIQUITOUS SENSOR NETWORK AND ITS NODE

Fig. 1 presents a brief description of the proposed IUSN and a functional block diagram of a sub-array node. In all, 16 microphone inputs are digitized with A/D converters; the sound information is stored in SRAM. They are then used for sound source localization and sound source separation. The power manager and voice activity detection (VAD) module deactivate the sound processing unit to conserve power: the sound processing unit is turned off if no sound exists around the microphone array. That power management is necessary because numerous microphones waste much power when not used.

Fig. 2 depicts 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 mutually connected to support their communication. Therefore, the sound gained by each node can be gathered to improve the sound source's SNR further. The system can be characterized as a huge microphone array by cooperating with surrounding nodes. Computations can be distributed among nodes. The system provides scalability in terms of

the number of microphones. Each node preprocesses acquired sound data. Then only compressed data—localized and separated sound—are communicated.

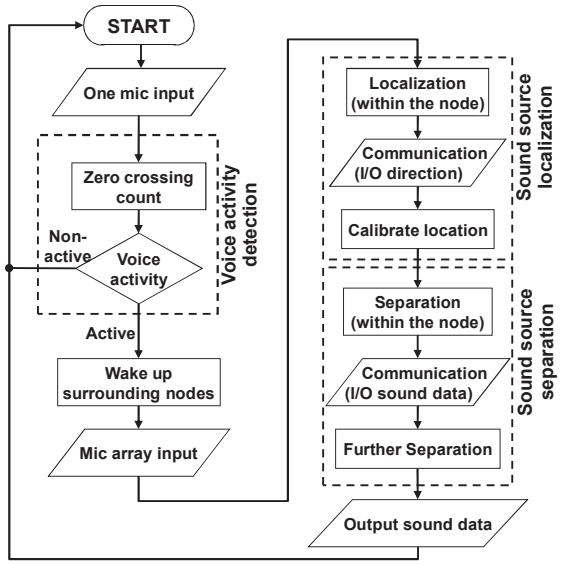

Fig. 2 Flow chart of an intelligent ubiquitous sensor node.

Regarding practical design, we implemented the intelligent ubiquitous sensor node on an FPGA board (SZ410, Suzaku; Atmark Techno Inc.) and microphones (ECM-C10; Sony Corp.). Fig. 3 shows prototype system photographs.

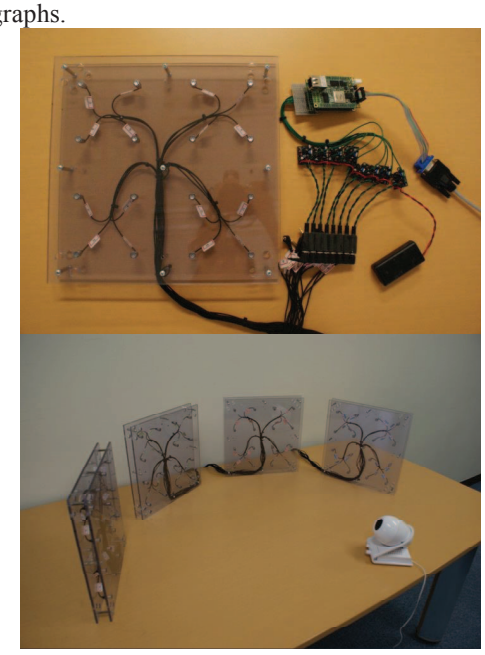

Fig. 3 System photographs: intelligent ubiquitous sensor node and a microphone array comprising sub-arrays.

We use low-power zero-crossing VAD, which is described in the next section. Sections 4 and 5 discuss the performance and accuracy of the sound source localization and sound source separation in our system, using simulated and measured data. For the system, gathering and processing localization data are important to improve the localization accuracy. Distributed localization data obtained with the multiple signal classification (MUSIC) algorithm [6] can be processed by a communication network in our system. Regarding the sound source separation, we use basic delay-and-sum beamforming both within a node and among nodes [7]. Therefore, the time accuracy between nodes strongly affects the final SNR of the sound source collected using the network.

3. VOICE ACTIVITY DETECTION

The microphone array network comprises numerous microphones, which would easily consume much power. Therefore, our intelligent ubiquitous sensor node must operate with a limited energy source and conserve power to the greatest extent possible. Sound processing that conserves power is effective because microphone amplifiers and the sound processing unit consume a certain amount of power even when they are sensing no sound.

In our previous work, we proposed a low-power VAD hardware implementation using a single microphone [8]. This custom hardware uses a zero-crossing algorithm for the VAD. Fig. 4 portrays the zero-crossing algorithm, as implemented on an FPGA in the ubiquitous sensor node as well.

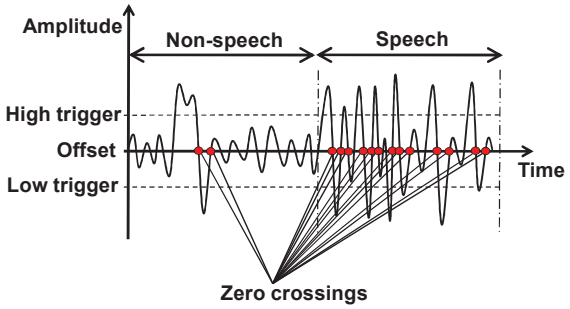

Fig. 4 Zero-crossing point example. The offset line shows the direct current (DC) component.

The zero crossing is the first intersection between an input signal and an offset line after the signal crosses a trigger line: the high trigger line or the low trigger line. Between a speech signal and non-speech signal, the appearance ratios of this zero crossing differ. The zero-crossing VAD detects this difference and outputs the beginning point and the end point of a speech segment. In our VAD algorithm, 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 2.1 mW is dissipated on the FPGA.

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 other 15 microphones are turned off. Furthermore, not all VAD modules in all nodes need operate: the system activates only some.

4. SOUND SOURCE LOCALIZATION

In this section, we propose a hierarchical localization method for sound source localization because the data communication bandwidth in our system is limited. We divide 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 [6] is chosen for node layer estimation because microphones on the node are limited to 16; the MUSIC algorithm can achieve higher resolution with fewer microphones. Since human speech includes wideband signals, we used a wideband MUSIC algorithm in the system. To find a relative direction, the sound source probability for P(θ) or P(θ, ϕ) is calculated for each node. Once the relative localization data are obtained, they are sent to a neighboring node to proceed to the next step.

We will localize the absolute sound source location in the network layer. A brief description of this method is presented in Fig. 5 with a two-dimensional coordinate of the sound source.

First, each node calculates P(θ) using the MUSIC algorithm and finds the maximum of P(θ) and corresponding θ. Then, the estimation data (a set of θs) are exchanged and the intersections for all combinations are calculated using their coordinates of the sub-arrays. At every intersection, a weighting ratio can be determined using P(θ)s. The coordinate of the sound source is estimated by finding the center of gravity of these weighting intersections.

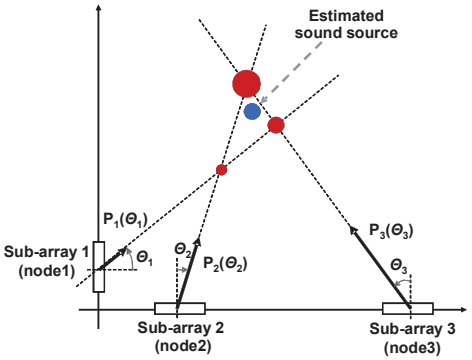

Fig. 5 Two-dimensional sound source localization.

We extend the two-dimensional method to three-dimensional localization, as shown in Fig. 6. First, the maximum P(θ, ϕ) and corresponding θ and ϕ are calculated on each node. We alternatively adopt the shortest line segment that connects two lines because we can usually find no exact intersection in three-dimensional space. We infer a point that divides the shortest line segment by the ratios of P(θ, ϕ)s as an intersection. The sound source is localized by calculating the center of gravity, as well, using the obtained intersections.

We verified the hierarchical localization through simulation, assuming an estimation result has a variation on every node. Fig. 7 portrays an example of the experiments. Fig. 8 shows the localization accuracy. The localization error is smaller when the arrays are numerous and the direction estimation is precise. Results show that minimization of the direction error increases the localization accuracy effectively. However, the sub-array number is not so influential.

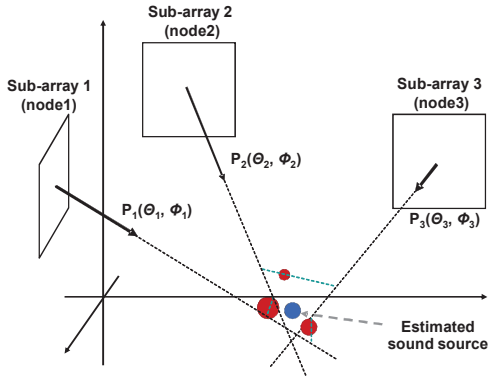

Fig. 6 Three-dimensional sound source localization.

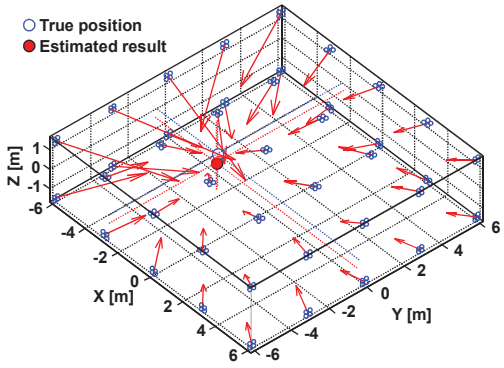

Fig. 7 Sound source localization experiment.

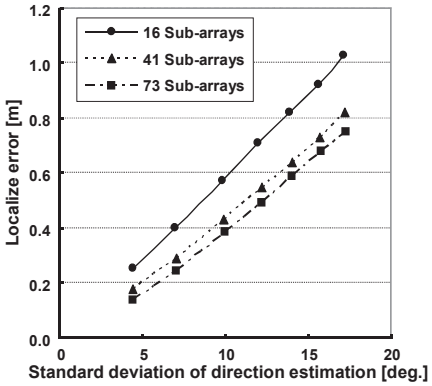

Fig. 8 Sound source localization accuracy.

5. SOUND SOURCE SEPARATION

Geometric techniques with position information and statistical techniques using no position information are two major sound source separation methods. The proposed system uses a geometric method, delay-and-sum beamforming, because the node positions are known. This method produces less distortion than statistical techniques; moreover, it requires few computations. For distributed

processing for sound source separation, it can be applied easily because it is based on summation (Fig. 9). Consequently, the key point for delay-and-sum beamforming for distributed nodes is the time difference among nodes.

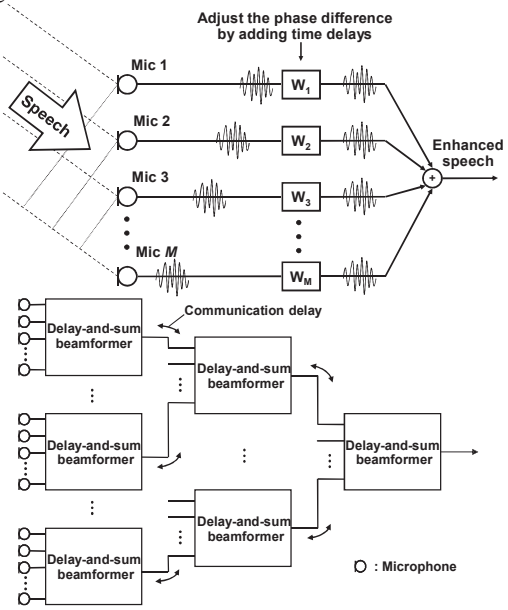

Fig. 9 Delay-and-sum beamforming with a node and among nodes.

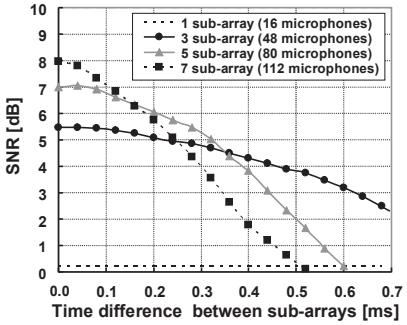

Fig. 10 SNR vs. time difference.

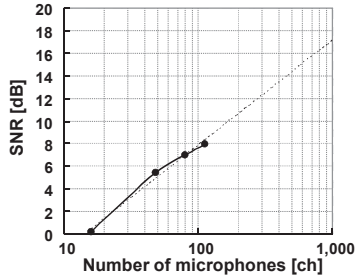

Fig. 11 SNR vs. number of microphones (time difference is 0.04 ms).

This experiment investigated the effects of the time difference among nodes. Fig. 10 shows the relation between the time difference and the SNR of the extracted speech; the SNR improves with the number of sub-arrays. However, the SNR decreases when the difference is too much. When many sub-arrays work together, the time difference between them should be 0.1 ms or less.

Fig. 11 shows that SNR improvement of 7.75 dB was gained with 112 microphones. We anticipate 15 dB or greater improvement using several tens of sub-arrays and several hundreds of microphones.

6. CONCLUSION

This proposed microphone array network realizes ubiquitous sound acquisition. A microphone array network comprising 16-microphone sub-arrays performed the following three operations within a node and with a network: 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. Low-power VAD was implemented to manage the nodes' power consumption. Thereby, the system achieves low power when speech is not active. The VAD module dissipates only 2.1 mW on an FPGA. Sound source localization is processed with the distributed nodes. The experimental result of the sound source separation demonstrated SNR improvement of 7.75 dB using 112 microphones. The system will achieve an SNR of 15 dB if the entire microphone network has more than several hundred 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/LCS Technical Memo MIT-LCS-TM-642*, Apr. 2004.

[2] Y. Tamai, S. Kagami, H. Mizoguchi, Y. Amemiya, K. Nagashima, T. Takano, "Real-time 2 dimensional sound source localization by 128-channel huge microphone array," *In Proc. of the IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN)*, pp.65-70, Sep. 2004.

[3] T. Wakabayashi, K. Takahashi, H. Iwakura, "Independent Component Analysis using Large Microphone Array," *Technical report of IEICE. EA 102(322)*, pp.29-34, Sep. 2002.

[4] Harvey F. Silverman, William R. Patterson III, James L. Flanagan, "The Huge Microphone Array", *IEEE Concurrency Volume 6, Issue 4*, pp.36-46, Oct.-Dec. 1998.

[5] H. F. Silverman, W. R. Patterson III, J. L. Flanagan, "The Huge Microphone Array, Part 2", *IEEE Concurrency Volume 7, Issue 1*, pp.32–47, Jan. 1999.

[6] R.O. Schmidt, "Multiple emitter location and signal parameter estimation," *In Proc. of the RADC Spectrum Estimation Workshop*, pp.243-248, Oct. 1979.

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

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