Processor Coupling Architecture for Aggressive Voltage Scaling on Multicores

Koji Kugata, Shinpei Soda, Yohei Nakata, Shunsuke Okumura, Shintaro Izumi, Masahiko Yoshimoto, and Hiroshi Kawaguchi

Graduate School of System Informatics
Kobe University
Kobe
Japan
657-8501
kugata@cs28.cs.kobe-u.ac.jp

Abstract: We propose novel multicore architecture in which dual processors (positive-true processor element and negativetrue processor element: pPE and nPE) can be combined and serve as a low-voltage/high-performance single PE. The coupling processor occupies a double area, but it operates at a lower voltage or faster than the original processor. The proposed scheme is suitable to voltage scaling on multicores that have abundant PEs. To evaluate and demonstrate the proposed architecture, we designed octa-core coupling DSP in a 65-nm CMOS technology, on which we confirmed a 1-MHz operation at a single supply voltage of 0.5 V.

1 INTRODUCTION

In multicore architecture, there are abundant processor elements (PEs). Several methods using the multicores have been proposed. For example, IBM's Cell processor increases its yield from 20% to 40% by disabling one of eight synergistic processor elements [1-2]. In another instance, a speed boosting technique, which called "Turbo Boost", is exploited in an Intel's Core i7 processor [3] and others. To dynamically vary the operating frequency, the Turbo Boost technology controls the number of active cores because the core frequency is determined by core temperature.

In this paper, we propose a power reduction technique using the abundant PEs. For a low-load task, not all PEs are needed; low power is desired rather than low processor usage. Under that condition, we can exploit an extra hardware resource to save the power. Aggressively reducing a minimum operating voltage (Vmin) with two PEs is allowed, which is effective for low power.

Fig. 1 shows the proposed multicores; there are a set of dual processors in which a positive-true PE (pPE) and negative-true PE (nPE) are adjacent. A normal task is carried out either pPE or nPE at a nominal voltage. For a low-load task, it is assigned to a

coupling processor (a pair of pPE and nPE) that can operate at a low voltage (voltage scaling). We call this "processor coupling architecture", which can also allocate a high-load task to a coupling processor running at a high frequency (speed boosting).

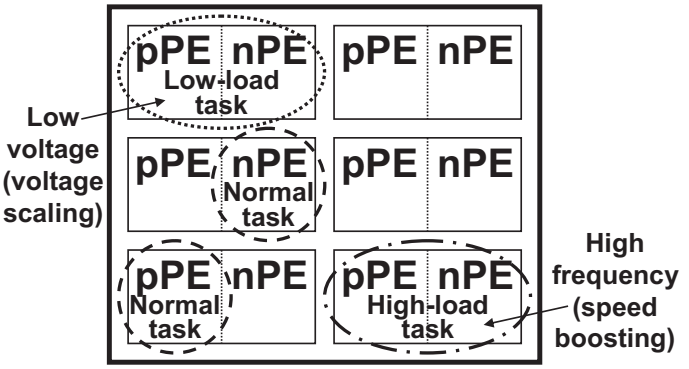


Figure 1: Proposed multicores: pairs of pPE and nPE.

2 PROCESSOR COUPING ARCHITECTURE

Fig. 2 portrays the two types (voltage scaling and speed boosting) of formations in the processor coupling architecture. Either of positive-true or negative-true path in logic circuits is selected depending on a process variation. Both SRAMs in the PEs are, however, used because their capacities are halved for the low-voltage (we will further explain this in detail in Subsections 2.1) or high frequency operation. In voltage scaling, coupling flip-flops ensure the low-voltage operation.

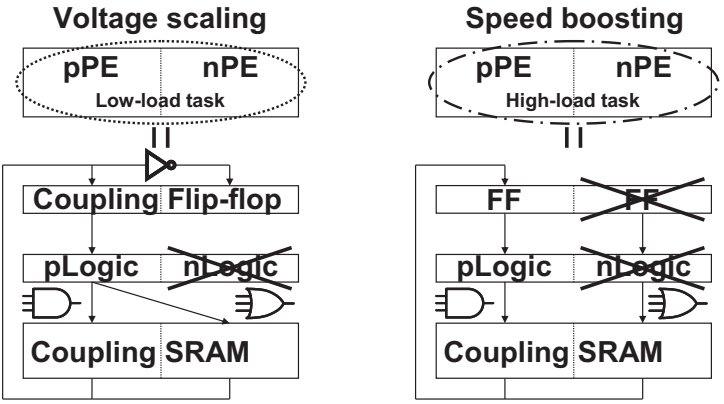


Figure 2: Processor coupling architecture: cases that a positive-true path is selected.

2.1 Coupling SRAM

Fig. 3 depicts coupling SRAM bitcells with a single-port (7T/14T) and dual-port (9T/18T) [4-5]. Two pMOSes are added to internal nodes in a pair of the conventional 6T and 8T bitcells. The coupling SRAM have two modes:

- Normal mode (7T and 9T): The additional transistors are turned off (/CL = "H"). The 7T bitcell acts as the conventional 6T cell.
- Coupling mode (14T and 18T): the additional transistors are turned on (/CL = "L"). Then, the internal nodes are shared by the memory cell pair. By doing so, a larger static noise margin can be obtained when either of the wordlines is merely activated, because a β ratio (a size ratio of a drive transistor to an access transistor) is doubled. When the both wordlines (WL0 and WL1) are activated, a high-speed operation is possible because a cell current is doubled (Fig. 4).

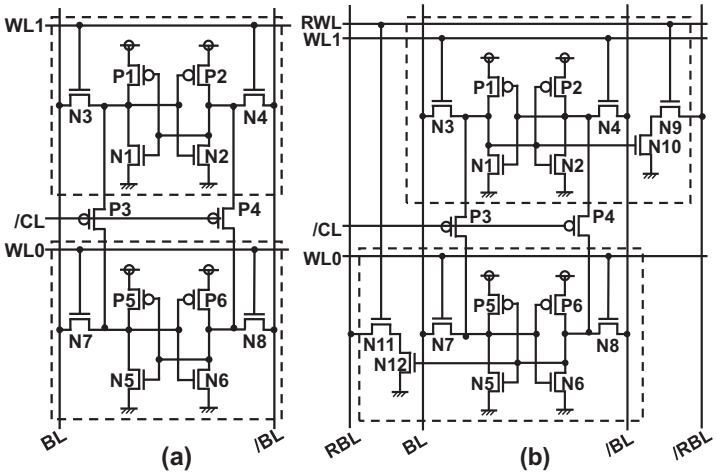


Figure 3: (a) 7T/14T and (b) 9T/18T coupling SRAM.

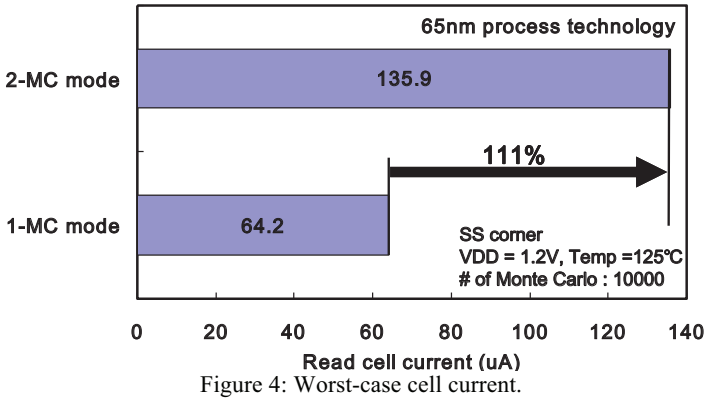


Figure 4: Worst-case cell current.

2.2 Coupling flip-flop

Fig. 5 shows a schematic of a coupling flip-flop pair. An unbalanced FS[1] (nMOS = fast and pMOS = slow) corner is the critical process corner in a low-voltage operation because of latch's data retention. This is similar to the worst retention condition in the 6T SRAM. As shown in Fig. 6, this coupling flip-flop can retain a datum at a voltage of 460 mV in the coupling mode. In the proposed flip-flops, the internal nodes are connected using four nMOS transfer gates, at which the two flip-flops complement the datum each other. The appended four gates are adaptively switched by a control signal (CTRL) according to the operating mode; the two flip-flops can independently operate by turning off the connecting gates in the normal mode.

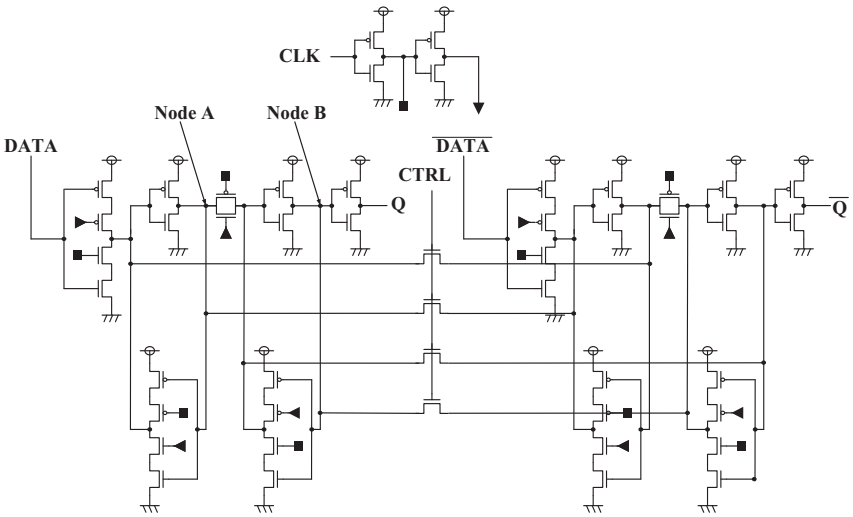


Figure 5: Coupling flip-flops.

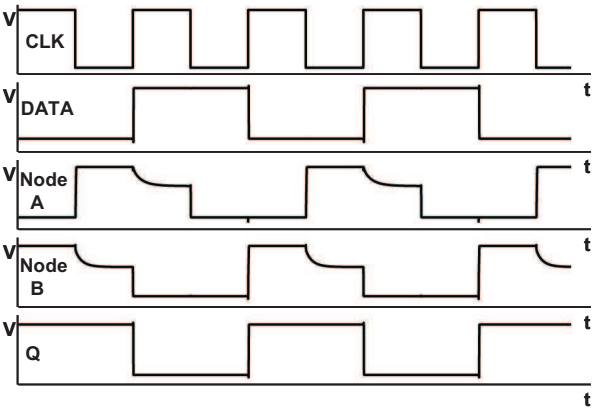


Figure 6: Simulated waveforms in the coupling flip-flops (1 MHz, 460 mV, FS corner).

[1] NMOS and pMOS out of the total variation in chip manufacturing VLSI, called process variation.

We designed and fabricated a coupling flip-flop on a test chip in a 65-nm process technology for measurement and verification. To evaluate behaviors at various process corners, triple-well structure is used for body biasing. In other words, applying body biases to pMOS and nMOS transistors give global Vt variations, which means that we can estimate circuit reliability under the global Vt variations. To guarantee the Vt control accuracy, we implemented a pMOS and nMOS test transistors on the chip for characteristic measurements.

Table 1 presents the body bias settings, at which four process corners (FF, FS, SF, and SS) are emulated, and measured Vmin's in the coupling flip-flop. In the table, ΔVtn (Δ|Vtp|) represents an nMOS (pMOS) transistor's threshold voltage difference from the fabricated CC transistor. Fig. 7 also portrays the measured Vmin when body biasing is applied. From these results, the Vmin of the flip-flop are reduced in the coupling formation at each process corner. This is because the coupling flip-flop improves the data retention characteristic. The voltage scaling mode can be applied to extremely low-power applications; for instance, biomedical sensing, sensor networking, and wearable computing.

Table 1. Body bias settings and measured Vmin in the coupling flip-flops.

Corner	ΔVPW [mV]	ΔVNW [mV]	ΔVtn	Δ|Vtp|	Vmin [mV]	
					Normal	Coupling
CC	±0	±0	±0	±0	384	166
FF	+500	-300	-146	-92	430	260
FS	+450	+400	-97	+67	540	424
SF	-600	-250	+74	-61	434	336
SS	-750	+600	+108	+99	430	210

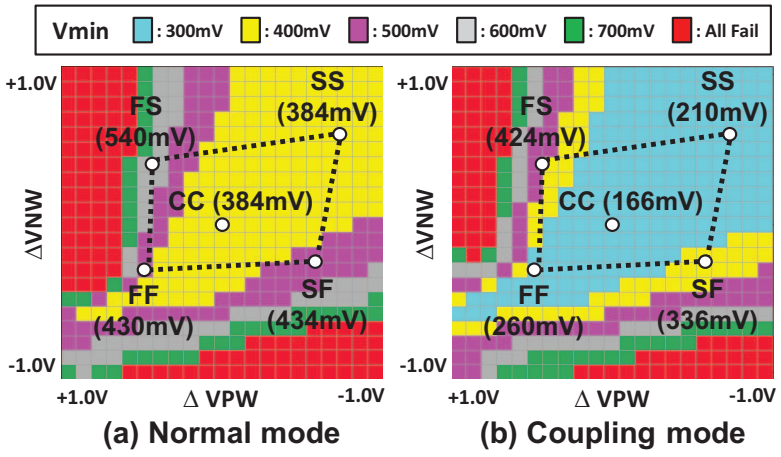


Figure 7: Measured Vmin in the coupling flip-flops at process corners.

2.3 Dual logic circuits

Usually, we consider the SS, TT, and FF (slow, typical, and fast) corners for logic synthesis. As described in the previous subsection, the worst process corner is, however, the unbalanced (FS or FS) corner at the very low supply voltage because the storage circuits like the flip-flop and SRAM are very sensitive to the unbalanced corner.

Fig. 8 illustrates the relationship between the process corner and eligible logic gate. For example, a decoder circuit in Fig. 9 (a) has a long falling time around the SF corner (Fig. 10 (a)) because a 3-input NAND has three stacked nMOS. On the other hand, an NOR logic has a long rising time around the opposite FS corner (Fig. 10 (b)). In our proposed scheme, because all data paths are dual, a better one can be selected according to process variation.

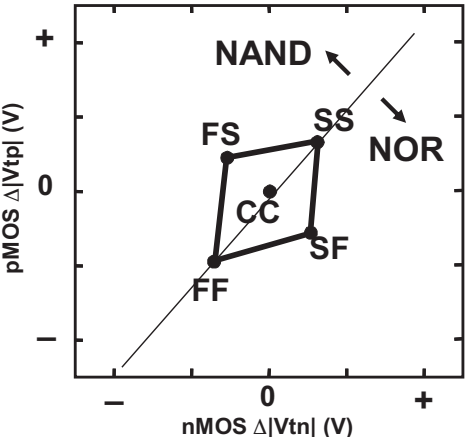


Figure 8: Process corner and eligible logic gate.

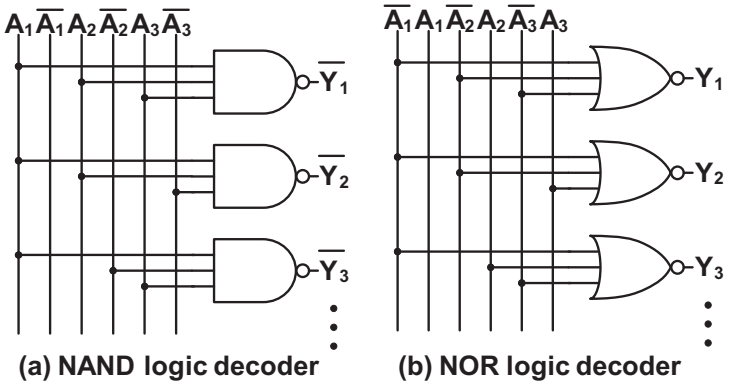


Figure 9: Decoder circuits (3-input NAND and NOR).

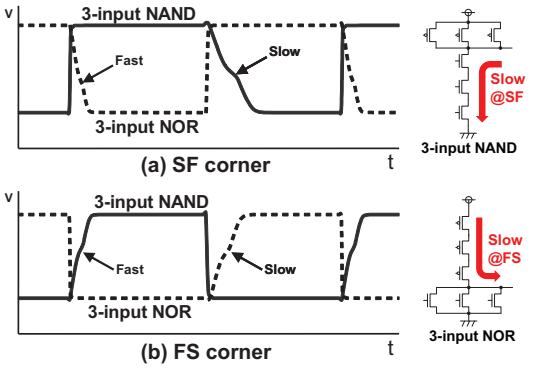


Figure 10: Rising and falling time comparison.

3 APPLICATION EXAMPLES

As an application example of the proposed coupling architecture, we designed octa-core digital signal processor (DSP) for a sound processing unit of microphone array networks.

3.1 Microphone array networks

Recent improvements in information processing technology have produced real-time sound-processing systems using microphone arrays [6]. The microphone array processes signal recordings and also performs noise reduction, sound source separation, speech recognition, speaker identification, and other tasks. To implement a microphone array as a realistic ubiquitous sound acquisition system with scalability, division of the huge array into sub-arrays with a multi-hop network is effective; an intelligent microphone array network was proposed in our previous work [7].

3.2 Coupling DSP multicores

Fig. 11 presents a block diagram of coupling DSP multicores for multi-channel signal processing. To perform SIMD operation for 16-channel sound processing, the proposed DSP consists of three units: instruction issuing unit (called master), sound processing unit, and network processing unit used to control packet and sound data communication.

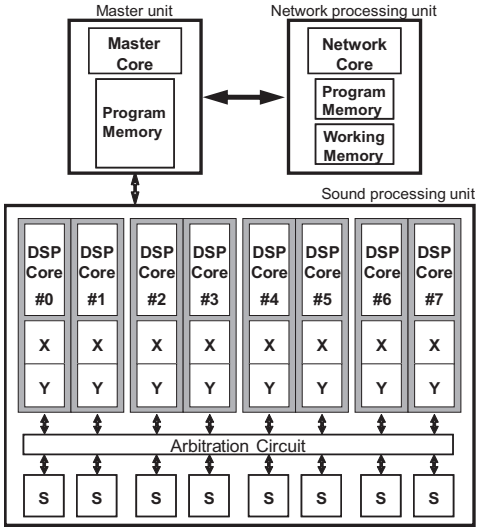


Figure 11: Proposed octa-core DSP multicores.

The master unit has a program memory (24 bits × 2048 word) comprising of the 7T/14T SRAM, decoder circuits, a program counter, and stack registers; it provides a same instruction to eight DSP cores in the sound processing unit. The four coupling DSP cores can be formed by the eight normal DSP cores. Each instruction is executed in six cycles (Fig. 12). The interrupt management and state transition are handled by the master unit.

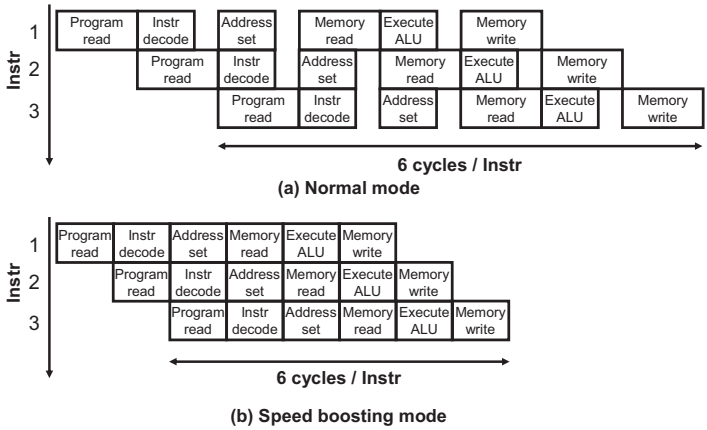


Figure 12: Six-stage pipelined execution.

As shown in Fig. 13, a DSP core has two working memories (X and Y in Fig. 11: each memory has 16 bit × 512 words) to process two-channel sound inputs. The shared-memories (S in Fig. 11) can be accessed from all the DSP cores through arbitration circuits for data exchanging. The X, Y, and shared memories can be accessed at the same time; the MAC and other operations can be executed in one cycle. The X, Y, and shared

memories consist of the 9T/18T dual-port coupling SRAMs to incorporate the dual-port feature for simultaneous read and write.

The working registers (40 bits) are implemented as an accumulator, which is comprised of the coupling flip-flops. As well, the other flip-flops in the DSP multicores consist of the coupling flip-flop.

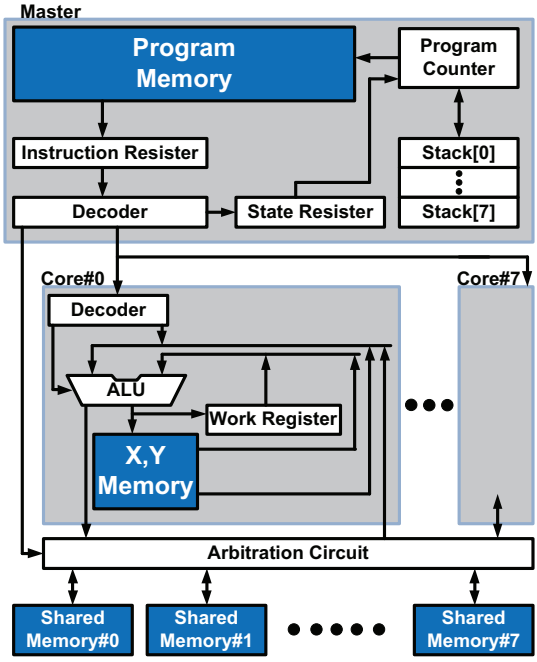
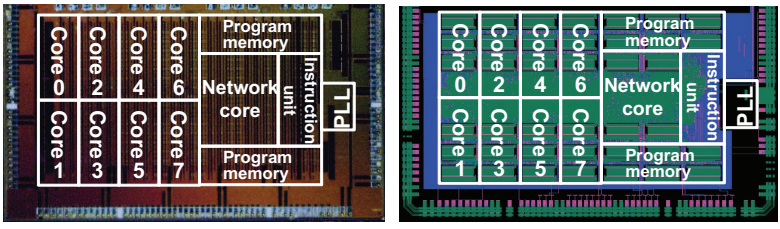


Figure 13: DSP core architecture.

3.3 VLSI implementation

We designed the octa-core coupling DSPs in the 65-nm CMOS process. Fig. 14 depicts the chip micrograph and layout plot image. The core size is 1722 × 2841 um^2. The power in the coupling mode (0.5 V, 1 MHz) is 0.73 mW.



(a) Chip micrograph (b) Layout

Figure 14: The coupling DSP multicores.

Summary

We proposed the processor coupling architecture for voltage scaling and speed boosting. The coupling SRAM achieves the low-voltage/high-speed operations by connecting two bitcells. The coupling flip-flop can run below 0.5 V at any process corner, which enlarges an operating voltage region in random logic circuits. The coupling logic circuits adapt to process variation by selecting an eligible data path. To evaluate the proposed coupling architecture, we designed octa-core coupling DSPs for sound signal processing in the microphone array network application. We fabricated a test chip in a 65-nm triple-well process, and confirmed that, in the coupling mode, the proposed architecture achieves a Vmin of 0.5 V and a power of 0.73 mW at a 1-MHz operation.

Acknowledgment

This research was supported in part by the Semiconductor Technology Academic Research Center (STARC) and KAKENHI (20360161).

References

[1] E. Sperling, “Turn Down the Heat... Please ― Interview with Tom Reeves of IBM,” EDN, July 2006.

[2] J. Kurzak, A. Buttari, P. Luszczek, and J. Dongarra, “The PlayStation 3 for High-Performance Scientific Computing,” Computing in Science and Engineering, pp.84-87, 2008. IEEE International Symposium on Quality Electronic Design (ISQED), pp. 98-102, Mar. 2008.

[3] J. Charles, P. Jassi, N. S. Ananth, A. Sadat and A. Fedorova, “Evaluation of the Intel Core i7 Turbo Boost feature,” In Proc. of IEEE International Symposium on Workload Characterization (IISWC), pp. 188-197, Oct. 2009.

[4] H. Fujiwara, S. Okumura, Y. Iguchi, H. Noguchi, Y. Morita, H. Kawaguchi, and M. Yoshimoto, “Quality of a Bit (QoB): A New Concept in Dependable SRAM,” In Proc. of IEEE International Symposium on Quality Electronic Design (ISQED), pp. 98-102, Mar. 2008.

[5] H. Noguchi, S. Okumura, T. Takagi, K. Kugata, M. Yoshimoto and H. Kawaguchi, “0.45-V Operating Vt-Variation Tolerant 9T/18T Dual-Port SRAM,” In Proc. of IEEE International Symposium on Quality Electronic Design (ISQED), pp. 219-222, Mar. 2011.

[6] M. Brandstein and D. Ward, “Microphone Arrays: Signal Processing Techniques and Applications,” Springer, 2001.

[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, 2010.