A SINGLE-CHIP SPEECH DIALOGUE MODULE AND ITS EVALUATION ON A PERSONAL ROBOT, PAPERO-MINI

Miki Sato, Toru Iwasawa, Akihiko Sugiyama, Toshihiro Nishizawa, Yosuke Takano

NEC Co mmon Platform Software Research Laboratories
Kawasaki 211-8666, JAPAN

ABSTRACT

This paper presents a single-chip speech dialogue module and its evaluation on a personal robot. This module is implemented on an application processor that was developed primarily for mobile phones to provide a compact size, low power-consumption, and low cost. It performs speech recognition with preprocessing functions such as direction-of-arrival (DOA) estimation, noise cancellation, beamforming with an array of microphones, and echo cancellation. Text-to-speech (TTS) conversion is also equipped with. Evaluation results obtained on a new personal robot, PaPeRo-mini, which is a scale-down version of PaPeRo, demonstrate an 85% correct rate in DOA estimation, and as much as 54% and 30% higher speech recognition rates in noisy environments and during robot utterances, respectively. These results are shown to be comparable to those obtained by PaPeRo.

***Index Terms*—** speech recognition, DOA estimation, noise cancellation, microphone array, echo cancellation, speech dialogue module

1. INTRODUCTION

Speech dialogue systems have been receiving particular attentions as a user interface for a wide variety of interactive applications, such as robots and car navigation systems. These applications are generally controlled by voice co mmands from a distance. A given co mmand is processed by a speech recognition system to generate a corresponding operation. It is also necessary to transform text information into an audible form by using a text-to-speech (TTS) conversion system. However, it is still challenging to perform off-microphone speech recognition, where the microphone is placed at a distance from the talker [1]. The target signal is seriously interfered by other signals and the ambient noise in noisy environments. Therefore, noise robustness is essential to speech recognition systems in the real environment.

To reduce undesirable influence by the ambient noise and the interference, signal-processing functions have been used for preprocessing the noisy speech. Among these functions are estimation of the direction of arrival (DOA) [2, 3], noise cancellation [4], beamforming with a microphone array [5], and echo cancellation [6]. DOA estimation identifies the direction of the voice co mmand so that the microphone directivity is steered towards the speech source. An adaptive noise canceller (ANC) and a microphone array (MA) reduce undesirable influence which cannot be sufficiently offset by the directional microphone. An acoustic echo canceller (AEC) suppresses an echo that is a part of robot speech leaking in the microphone signal and contaminating the voice co mmand.

In robot applications, these functions are generally implemented by software on a platform based on a personal computer (PC) [7]. It is sometimes necessary to share computational power with other applications on the same platform. Considering that a larger number of complex applications are required on a robot, it is desirable to have a speech dialogue module on a separate platform so that the PC-based platform can be fully devoted to more complex and computationally intense applications on the robot. On such a separate module, the performance of the speech dialogue functions becomes more stable and guaranteed. In addition, a compact speech dialogue module helps promote human-robot interactions, with its portability, on more robots that otherwise would not have such an interface.

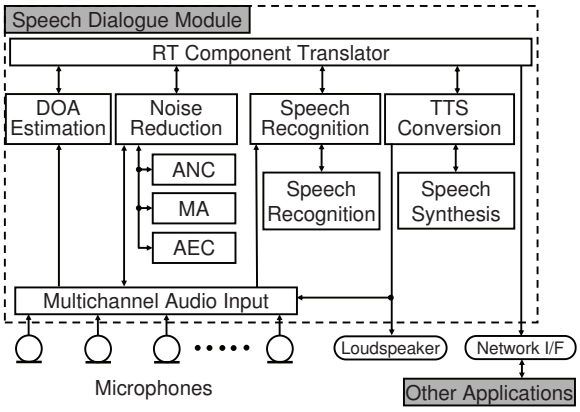

Fig. 1. Block diagram of Speech Dialogue Module.

This paper presents a compact speech dialogue module and its evaluation on a personal robot. This module offers dialogue functions similar to a personal robot PaPeRo [8] on an application processor that was developed primarily for mobile phones to provide a compact size, low power-consumption, and low cost. In the following section, functions of the speech dialogue module are described with the hardware for their implementation. Section 3 presents evaluation results of a near-field DOA estimator, a noise-robust ANC with variable stepsizes, an adaptive beamformer for MA, and a noise-robust AEC in the real environment.

2. SPEECH DIALOGUE MODULE

2.1. Speech Dialogue Functions

A block diagram of the speech dialogue module is illustrated in Fig. 1. This module consists of speech recognition (word recognition), DOA estimation, noise reduction, and TTS conversion as speech dialogue functions. Noise reduction has three subfunctions, namely, an adaptive noise canceller (ANC), a microphone-array (MA) beamformer, and an acoustic echo canceller (AEC). They operate separately and a desired output is manually selected. These functions

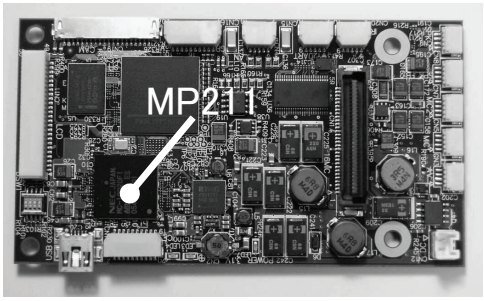

Fig. 2. Speech Dialogue Module.

Table 1. Specifications of the Speech Dialogue Module

Item	Specification
CPU	ARM9(192 MHz) × 3 DSP(SPXK6 192 MHz) × 1
Memory	128MB +128MB (w/ extended board) Flash 64MB
Audio Interface	microphone inputs 2ch × 2 +16ch (w/ audio board) speaker output 2ch
Image Interface	camera input × 2 video output, LCD output
Other Interface	USB, LAN, IrDA, GPIO CF Card (w/ extended board)
Size	55 mm × 100 mm × 32 mm (w/ audio and extended boards)

work as RT (Robot Technology) components for RT middleware [9], and can be controlled by other network-connected applications.

2.2. Hardware

Figure 2 shows a picture of the speech dialogue module whose specifications are illustrated in Table 1. An application processor, MP211 [10], primarily designed for mobile phones, is employed for a sufficient processing power. It consists of one DSP and three ARM9 cores and runs on a Linux operating system. For audio input interface, this module is equipped with synchronous microphone inputs on an extended board that are extensible to 16 channels, as well as 2 types of 2-channel synchronous microphone inputs on the main board. In addition, the module also has some peripheral interfaces such as 2-channel loudspeaker outputs, 2-channel camera inputs, an LCD output, a USB and a LAN interfaces. It is possible to use a compact flash memory (CF) card on an extended board.

2.3. Implementation of the functions

The functions of the module were distributed to an ARM9 and a DSP cores running at 192 MHz. The task distribution between the ARM9 and the DSP are depicted in Fig. 3. Speech recognition, TTS conversion, and RT component translator operates on the ARM9. DOA estimation and noise reduction are decomposed into core-functions and control blocks. The noise-reduction core consists of three sub-cores, namely, ANC core, MA core, and AEC core. The control blocks operate on the ARM core and the core-fuction blocks on the DSP core. The input signals are converted into a digital form at a rate of 11025 Hz and saved in multi-ring buffers on an internal memory

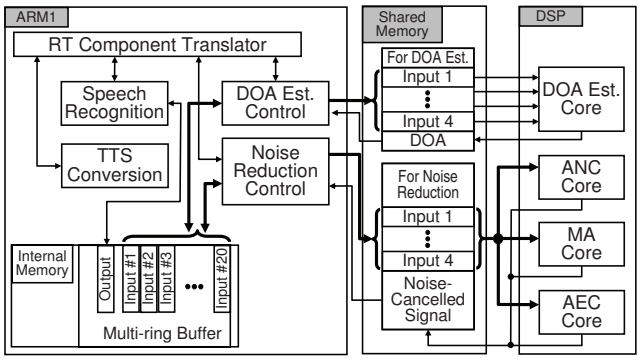

Fig. 3. Task distribution between ARM and DSP cores.

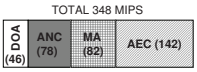

Fig. 4. Computational load for each function.

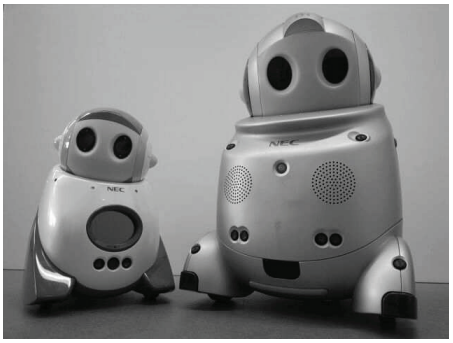

Fig. 5. PaPeRo-mini (Left) and PaPeRo (Right).

of the ARM. The computational load for each noise-reduction function is compared in Fig. 4.

3. EVALUATION

3.1. Platform: PaPeRo-mini

The speech dialogue module was installed in a PaPeRo-mini which is a scale-down version of PaPeRo, a partner robot based on a Windows PC. Figure 5 depicts PaPeRo and PaPeRo-mini whose specifications are compared in Table 2. PaPeRo-mini is an autonomous mobile robot with a size of 250 × 170 × 179 mm (HWD) and a weight of 2.5 Kg. Equispaced eight omnidirectional microphones are mounted around the neck and 2-channel loudspeakers are mounted in the bottom. It also has CCD cameras, ultrasonic sensors, infrared sensors, touch sensors, a pyroelectric sensor, and an LCD.

3.2. DOA (direction of arrival) Estimation [3]

Figure 6 (a) depicts the evaluation environment in a room with a background noise level of 40 dBA. One sentence spoken by 5 different males and females were presented 10 times at 75 dB from a loudspeaker at 1.0 m in elevation. PaPeRo-mini was placed 1.5 m away and turned with a step of 30 degrees to make 12 different DOAs. The microphone arrangement of PaPeRo-mini and PaPeRo are illustrated in Fig. 7.

Table 2. Specifications of PaPeRo-mini and PaPeRo

	PaPeRo-mini	PaPeRo
CPU	MP211	Pentium-M 1.6 GHz
OS	Linux	Windows XP
Audio Input	Omnidirectional Mic x8	Omnidirectional Mic x7 Directional Mic x1
Audio Output	Stereo Loudspeakers Line Output x2	Stereo Loudspeakers Line Output x2
Image Input	Stereo CCD Camera	Stereo CCD Camera
Image Output	Composite Video LCD	Composite Video RGB
Other I/F	IrDA USB	Remote Control USB
Battery	Li-ion 74Wh Operating Time 8h	Li-ion 60Wh Operating Time 2h
Size Weight	250x170x179 mm 2.5 kg	385x248x245 mm 5.0 kg

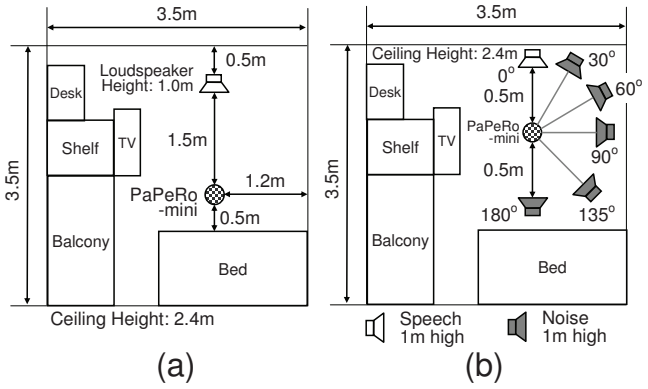

Fig. 6. Evaluation Environment. (a) DOA, (b) ANC/MA/AEC.

Figure 8 depicts the evaluation result. Any DOA estimation results other than those with insufficient power, correlation, or inconsistent DOAs among different microphone pairs are considered as detection. The correct answer has a margin for an error of ± 15 degrees from the true DOA. For comparison, the evaluation result of PaPeRo is depicted in Figure 8. Parameters for height adjustment [3] were selected as $d = 1.5$ m and $h = 1.0$ m for a typical robot use at home. The correct answer rate is slightly more degraded than others for 60 degrees. However, average rate of correct answers reaches 83% which is comparable to PaPeRo.

3.3. ANC (adaptive noise canceller) [4]

Speech recognition was performed with noise-cancelled speech by the ANC. The evaluation environment, prepared in the same room as that for DOA estimation, is depicted in Figure 6 (b). 450 utterances by 9 different males, females and children were presented at a distance of 1.0 m, a height of 1.0 m, and a level of 70 dB. A loudspeaker presenting a commercial TV-program was placed 1.0 m away as the noise source in a direction of 90, 135, or 180 degrees at a level of 60-65 dB. A dictionary of 50 recognition and noise-rejection words [11] was used for speech recognition. For signal input, a front-side and a rear-side microphones among the eight around the neck of PaPeRo-mini were used as the primary and the reference microphones.

Figure 9 demonstrates the speech recognition rate. For comparison, the average speech-recognition rate of PaPeRo [4], is depicted in Figure 9. It was evaluated on 1500 utterances by 30 different males, females and children at a distance of 0.5 m and 1.5 m at a level of 70 dB. The noise level was set at 55-60 and 65-70 dB. The recognition rate with an ANC is improved by more than 40% and the maximum improvement reaches 54% with noise arriving from behind the robot. When there is no noise, the recognition rate of PaPeRo-mini is more than 15% lower than that of PaPeRo. It comes from the microphones. PaPeRo uses a directional microphone, while PaPeRo-mini uses an omnidirectional microphone. However, due to the ANC, the recognition rates in noisy environment are almost comparable.

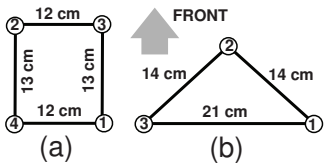

Fig. 7. Microphone Arrangement (Top View, Distance in cm). (a) PaPeRo-mini, (b) PaPeRo

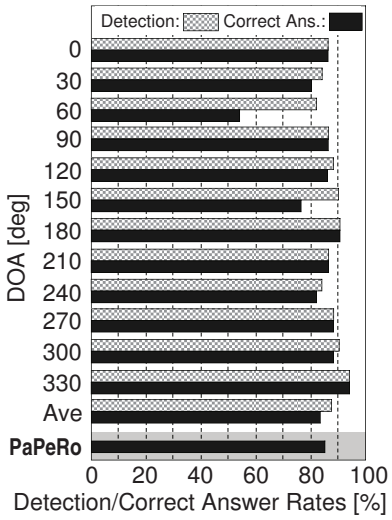

Fig. 8. DOA Estimation Result.

3.4. MA (microphone array) [5]

In the case of MA, the conditions for evaluation were same as those for the ANC except the noise directions. The noise source was placed in a direction of 30, 60, or 90 degrees. For the MA, four microphones arranged linearly with 0.02 m spacings were mounted on the front-side of PaPeRo-mini. Figure 10 depicts the speech recognition rate in comparison with an average rate by PaPeRo in the same condition as that for PaPeRo-mini. Due to the MA, the recognition rate is improved by more than 20% and the maximum improvement reaches 40% with noise arriving from the front of the robot. The recognition rate of PaPeRo-mini with the MA is comparable to that of PaPeRo.

3.5. AEC (acoustic echo canceller) [6]

Speech recognition was performed with echo-cancelled speech by AEC. The condition of evaluation was the same as that for the ANC

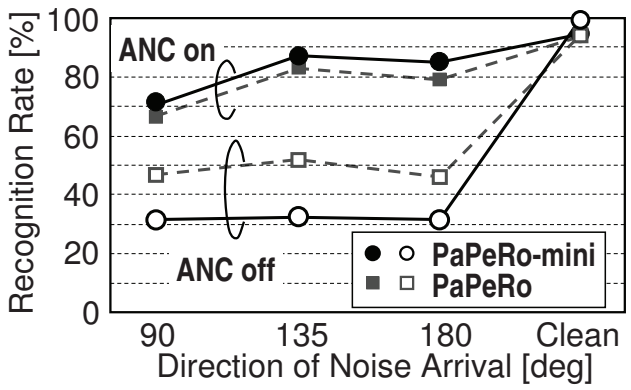

Fig. 9. Speech Recognition Result (ANC).

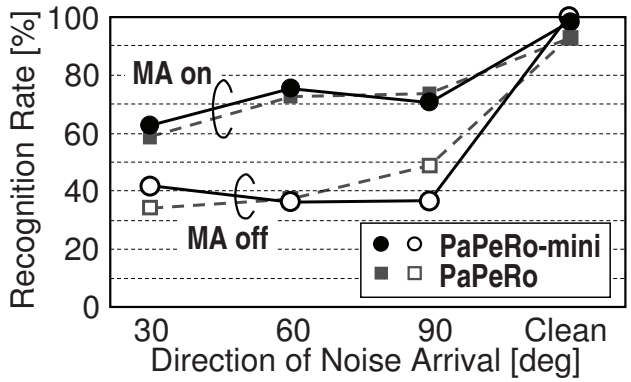

Fig. 10. Speech Recognition Result (MA).

and the MA except that there was no noise source. The music source was presented as an echo at 60-65 dB from a loudspeaker mounted on the bottom of PaPeRo-mini. A microphone same as the primary microphone for the ANC was used to capture the echo and the target speech. Figure 11 depicts the speech recognition rate in comparison with the PaPeRo data in the same environment. The echo level for PaPeRo was at 55-60 and 65-70 dB. Due to the AEC, the recognition rate is improved by 30% with the sound from the loudspeaker of the robot. The speech recognition rate by PaPeRo-mini was equivalent to that of PaPeRo.

4. CONCLUSION

A single-chip speech dialogue module and its evaluation on a personal robot has been presented. This module has been implemented on a single-chip application processor to provide a compact size, low power-consumption, and portability. It has been equipped with direction of arrival (DOA) estimation, adaptive noise cancellation (ANC), a microphone array (MA) beamforming, and acoustic echo cancellation (AEC) for speech recognition in noisy environment. Evaluation results obtained on PaPeRo-mini in real environment have demonstrated an 85% correct rate in DOA estimation, and as much as 54% and 30% higher speech recognition rates in noisy environments and during robot utterances, respectively.

5. ACKNOWLEDGMENT

This development was supported in part by a common platform development project for next-generation robots of NEDO (New Energy and Industrial Technology Development Organization).

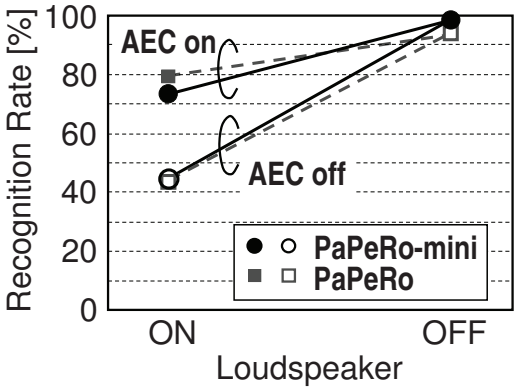

Fig. 11. Speech Recognition Result (AEC).

6. REFERENCES

[1] H.-G. Hirsch and D. Pearce, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions, " Proc. ISCA ITRW Asr 2000, Sep. 2000.

[2] Y. Kaneda, "Sound source localization; Robot audition system from the signal processing point of view, " Proc. 22nd AI Challenges, Vol.22, pp.1–8, Oct. 2005 (in Japanese).

[3] M. Sato, A. Sugiyama, O. Hoshuyama, N. Yamashita, and Y. Fujita, "Near-Field Sound-Source Localization Based on a Signed Binary Code," IEICE Trans. Fundamentals, Vol.E88-A, No.8, pp.2078–2086, Aug. 2005.

[4] M. Sato, A. Sugiyama, S. Ohnaka, "An adaptive noise canceller with low signal-distortion based on variable stepsize subfilters for human-robot communication, " IEICE Trans. Fundamentals, Vol.E88-A, No.8, pp.2055–2061, Aug. 2005.

[5] O. Hoshuyama, A. Sugiyama, A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters, " IEEE Transactions on Signal Processing, Vol.47, pp.2677–2684, Oct. 1999.

[6] A. Sugiyama, J. Berclaz, and M. Sato, "Noise-robust double-talk detection based on normalized cross correlation and a noise offset, " Proc. ICASSP2005, pp.153–156, Mar. 2005.

[7] M. Sato, A. Sugiyama, O. Hoshuyama, N. Yamashita, S. Ohnaka, and Y. Fujita, "The voice interface of Personal Robot, PaPeRo," J. of Acoust. Soc. of Japan, Vol. 62, No. 3, pp.1–9, Mar. 2006 (in Japanese).

[8] Y. Fujita, "Personal Robot PaPeRo," J. of Robotics and Mechatronics, Vol.14, No.1, pp.60-63, Jan. 2002.

[9] N. Ando, T. Suehiro, K. Kitagaki, T. Kotoku, Y. Woo-Kuen, "RT-middleware: distributed component middleware for RT (robot technology)," Proc. IROS 2005, pp.3933-3938, Aug. 2005.

[10] S. Torii et al., "A 600MIPS 120mW 70μA Leakage Triple-CPU Mobile Application Processor Chip," Proc. of ISSCC2005, 7.5, Feb. 2005.

[11] T. Iwasawa, "Speech Recognition Interface for Personal Robot "PaPeRo," Proc. 13th AI Challenges, Vol.13, pp.17-23, Jun. 2001.