Network-based multi-channel signal processing using the precision time protocol

Yoshifumi Chisaki*, Dan Murakami and Tsuyoshi Usagawa†
* Kumamoto University, Kumamoto, Japan
E-mail: chisaki@cs.kumamoto-u.ac.jp
† Kumamoto University, Kumamoto, Japan
E-mail: tuie@cs.kumamoto-u.ac.jp

***Abstract*—A conventional microphone array system uses a conductor to wire from a microphone to an input via an amplifier. While, a wireless transmission for an array system makes the configuration flexible, and it is expected to provide novel applications widely, such as measuring of impulse response in wide area. Not only in acoustic research field but also in research area of remote sensing of body motion and so on, data acquisition with a precise time synchronization is essential. When a signal received at distributed microphone's position is sent over a computer network, the data is packetized and can include some information at receiving point. When a time is included as information at data acquisition position, the time depends on each data acquisition client device. Since it is possible to use network time protocol or precision time protocol, multi-channel signal processing can be achieved with ease. This paper proposes multiple signals transmission system over a computer network with a time code embedding to synchronize those signals. The signal from a client is reconstructed at a server with the time code. Since the time differences between clients affects to performance of the multichannel signal processing, smaller error in time at a client is preferred. This paper discusses how the error in time between channels affects to performance of the distributed microphone array system.**

I. Introduction

Recent development of computer networking and portable devices contribute to wide area in life. Since portable devices including cellular phone equips one or two microphones and data connection using WiFi or 3G, acoustic signal can be transmitted with ease. Thus, the portable device has potential to work as a part of distributed acoustic signal acquisition system. The conventional microphone array system is widely used for many situations, such as TV conference, and requires complicated wiring. While, wireless signal acquisition introduces the flexibility for configuration. Moreover, data acquisition with a precise time synchronization is essential not only in acoustic research field but also in research area of remote sensing of body motion and so on.

The distributed acoustic signal acquisition system consists of clients and server. The data acquired with a client device at observation position is sent to a server which performs multi-channel signal processing. Each client device has analogue to digital converter and should have a clock for time synchronization. Synchronization with higher precision in time is essential to obtain better performance on multi-channel signal processing. In a computer network research field, network protocol time (NTP) is a well-known time synchronization method, and is implemented not only on a potable device but also on a personal computer. The method works with a millisecond order error in all the networks. While, the precision time protocol (PTP) can synchronize clients with smaller error. The PTP is designed for measurement and control systems and defined in IEEE 1588-2008. However, the PTP works only in local area network (LAN). The multi-channel signal processing which uses NTP discussed in the literature[1]. This paper examines performance by the multi-channel signal processing with PTP.

II. Measuring signal synchronization accuracy with PTP

Signal synchronization accuracy with PTP protocol is measured. Table I shows configuration of measurement. PTP uses specialized hardware to obtain precise clock and PTP softwares(client-server) usually. In case of rack of the specialized hardware, a precision of time adjustment is degraded because clock interval is insufficient on a general device. However, the specialized hardware is quite expensive. This paper focuses on performance by only the software of PTP.

Two terminals adjust their internal clocks with a PTP server, record the same white noise signal via line-in. The PTP server and recording terminals were placed in the same 1000BASE-T network segment in a laboratory student room. About 20 terminals are connected to the network.

After time alignment based on embedded timecode, time lags of signals were estimated by cross correlation of two recorded signals. Note that these time lags include not only errors in time of PTP synchronization, but also latencies of the recording systems. The number of trials is 5000 in a session, and each measurement is performed every 10 seconds.

Fig. 1 shows an example of time lags after time alignment with embedded time codes. An average of time lags was -0.07 ms and standard deviation of it was 0.13 ms. Since the average was 0.25 ms and the standard deviation was 0.44 ms in case of NTP in the previous study [1], the error in time is decreased by PTP. Although PTP works only in LAN, use of PTP is effective.

These measured time lags were added to generated signals in the simulations.

TABLE I
CONFIGURATION OF MEASUREMENT

Network	1000BASE-T full-duplex (In a laboratory)
PTP clients	MacBook
Polling interval	16 [s]
PTP daemon program	Ver. 2.2.0
Audio I/O	ONKYO SE-U55 (USB)
PTP server	Mac mini
PTP daemon program	Ver. 2.2.0

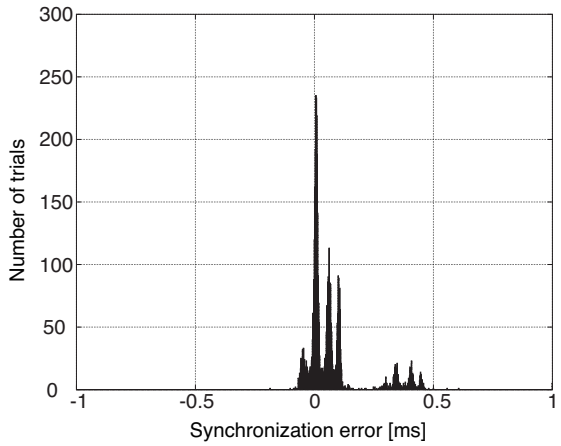

Fig. 1. Error in time.

III. METHOD

In this paper, a spectral filtering approach is proposed and simulations are performed. Fig. 2 illustrates a whole blocks of the proposed method. The method is built as n-input n-output system. The precision time protocol (PTP) is utilized to synchronize signals from each terminal.

Fig. 3 illustrates a block diagram of the proposed method. The diagram can be considered as a part of the whole system. Each speaker uses a PC terminal which has a microphone. Here in after, a microphone for a target speaker is called a target microphone and microphones for other speakers are called interference microphones.

Since most influential interference speakers to a target speaker are nearest neighbors, signals acquired at nearest two interference microphones are utilized to enhancement of the target signal.

For simple explanation, $\mathrm{SP_C}$ is assumed to a target speaker, $\mathrm{SP_L}$ and $\mathrm{SP_R}$ are assumed to interference speakers. Thus, a microphone $\mathrm{M_C}$ for the target speaker, $\mathrm{M_L}$ and $\mathrm{M_R}$ for each interference speaker.

A. FIFO buffer

Signal from terminals are transmitted over TCP/IP network individually. Since the characteristics of the network, each signal has individual time lag from recorded. Thus, signals are once stored to "FIFO buffer." Then, stored signals are passed to "Time alignment" block.

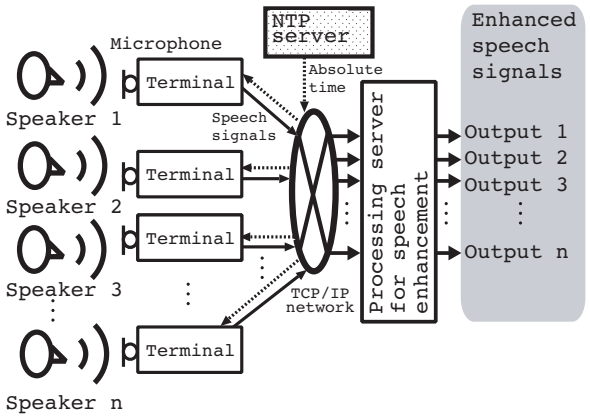

Fig. 2. Whole blocks of the proposed method.

B. Time alignment of signals acquired over TCP/IP network

At the beginning, internal clocks of all terminals are synchronized with the "PTP server." Then, signals observed at each terminal are transmitted to processing server over TCP/IP based network. Timecode is embedded to each signal to obtain time synchronization between all signals. Each time code is based on a internal clock of each terminal which is synchronized with the PTP server. Thus, time lags caused by internal clock of each terminal or the PTP server is included to each time code.

C. FFT

After the "Time alignment" is performed, signals are converted to time-frequency (T-F) domain by the Fast Fourier Transform (FFT) at "FFT" blocks. FFT divides a time-domain signal into a series of small overlapping frames. Each of these frames are windowed and transformed by Fourier transformation Transformed signals are denoted in T-F domain as $S_C(k,i), S_L(k,i)$ and $S_L(k,i)$ where k denotes frequency bin index and i denotes frame index.

D. Time-frequency mask estimation

The "Time-frequency mask estimation" blocks estimate masks applying weight to each T-F unit, such that spectro-temporal regions that are dominated by the target speech are emphasized, and regions that are dominated by interference speeches are suppressed. The T-F masks are estimated for each pair of spectrums, $S_C(k,i)$ and $S_L(k,i)$, $S_C(k,i)$ and $S_R(k,i)$ by comparing amplitude of each spectro-temporal region. For example, T-F regions that have higher amplitude in $S_C(k,i)$ than $S_L(k,i)$ are emphasized since these regions are assumed to belong $S_C(k,i)$, and vice versa.

This idea is based on the basic two input and two output binary masking method that assumes sparseness of a pair of speech signals [2]. In the proposed method, the idea is expanded to asynchronous three signals by assuming them as two pairs of signals.

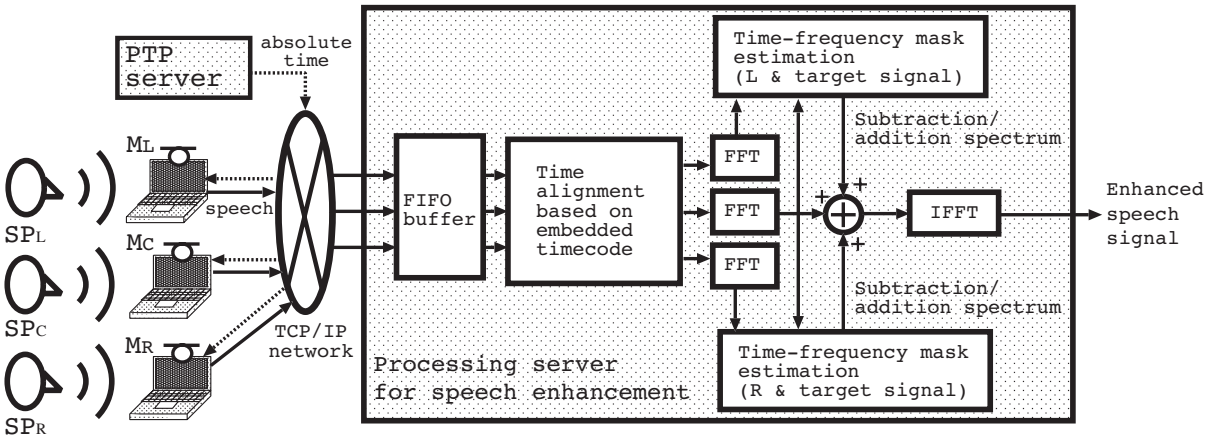

Fig. 3. Block diagram of the proposed method.

Amplitude ratio of each T-F component between $S_C(k,i)$ and $S_L(k,i)$ is given by

$$R_{CL}(k,i) = 10\log_{10}\frac{|S_C(k,i)|}{|S_L(k,i)|} \ \forall\ k,i, \tag{1}$$

where $\forall\ k,i$ means the equation is applied for all frequency bin index k and frame index i. Then, a T-F mask $M_{LC}(k,i)$ is given by

$$M_{LC}(k,i) = \begin{cases} 1 & if\ R_{CL}(k,i) > 0 \\ -1 & otherwise \end{cases} \ \forall\ k,i, \tag{2}$$

An addition/subtraction amplitude spectrum is given by applying the mask $M_{LC}(k,i)$ to $|S_L(k,i)|$ as following:

$$\hat{S}_{LC}(k,i) = |S_L(k,i)| \cdot M_{LC}(k,i)\ \forall\ k,i. \tag{3}$$

$\hat{S}_{RC}(k,i)$ is also acquired by all the same procedure as above that is processed between $S_C(k,i)$ and $S_R(k,i)$. The target utterance is emphasized by adding estimated two addition/subtraction amplitude spectrum as following:

$$S_{CLR}(k,i) = |S_C(k,i)| + \hat{S}_{LC}(k,i) + \hat{S}_{RC}(k,i)\ \forall\ k,i. \tag{4}$$

Phase spectrum of output signal is preserved as originally observed signal $S_C(k,i)$ as following:

$$\hat{S}_C(k,i) = S_{CLR}(k,i) \cdot e^{i\arg S_C(k,i)}\ \forall\ k,i. \tag{5}$$

Fig. 4 illustrates spectrograms of example signals in a case. In the Fig. (a), (b) and (c) are dry source of SP_L, SP_C and SP_L; (d), (e) and (f) are signals observed at M_L, M_C and M_L.

With the signals, mask filters to enhance the targe speech signal are estimated as illustrated in Fig. 5 by two pairs of signals, S_C and S_L, S_C and S_R. By compairing fig.4 (b), (e) and (g), the masks can be seen as a filter to suppress noise components from interferences and enhance the target speech.

E. IFFT

Output speech is calculated by "IFFT" using overlap-save method. By this process, the target signal expressed in T-F regions are converted to waveforms of time domain.

Fig. 6 illustrates waveforms of (a) a target speech, (b) a mixed signal observed at the target mic and (c) a mixed signal processed with the proposed method. According to it, envelope of waveform is certainly recovered like the target speech.

IV. Simulation

Simulations were performed in several settings as shown in Fig. 7 and Table II. Simulation I is performed in an anechoic room and simulation II is performed in laboratory student room. A distance between each loudspeaker and a microphone for it is 0.5 m, which is recommended distance of ITU-T P.340 for hands-free terminals [3].

A. Signal generation

Dry sources of speech signals are convolved with impulse responses to generate signals observed at microphones. Impulse responses were measured by time-stretched pulse radiated from each loudspeaker to each microphone [4].

B. Utterances used in the simulation

1000 set of utterances consist of 212 phoneme-balanced words are selected randomly from Tohoku University and Panasonic isolated spoken word database [5]. Each set includes 3 utterances to be radiated from loudspeakers simultaneously.

C. Evaluation method

Utterances are radiated from 3 speakers simultaneously. A radiated power ratio between loudspeakers, namely SNR of a target signal to an interference signal is set to 0 dB. The FFT is computed using a Hann window of 1024 samples of length, with 768 samples overlap.

Signals processed with the proposed method are evaluated by recognition rates of the ASR, Julius [6]. Each utterance is recognized as an isolated-word. Speaker-independent acoustical model provided with Julius dictation kit is used for speech

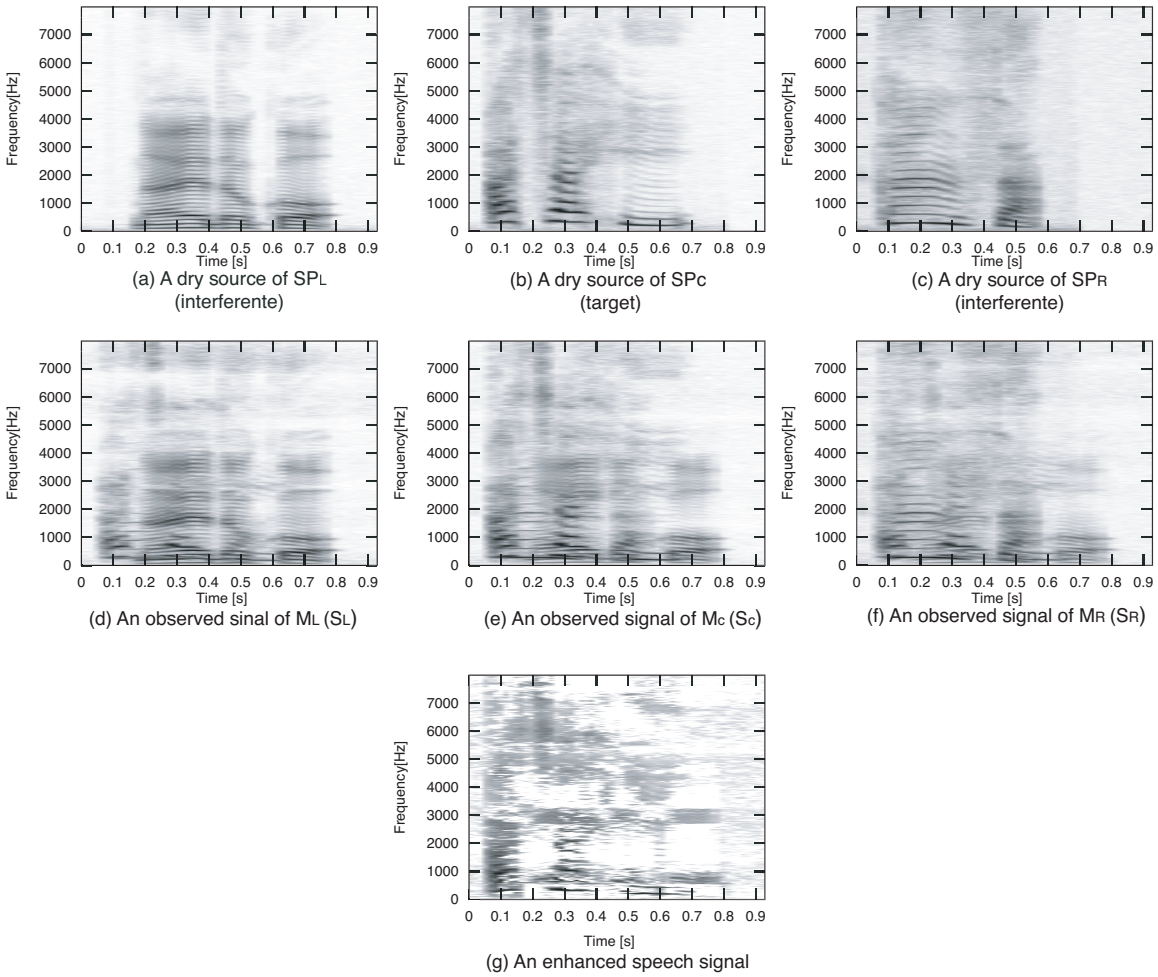

Fig. 4. An example of spectrograms of dry sources, observed signals and an enhanced speech signal.

TABLE II
CONDITIONS FOR SIMULATION

Simulation	I					II				
	(a)	(b)	(c)	(d)	(e)	(a)	(b)	(c)	(d)	(e)
Environment	Anechoic room					Laboratory student room				
Interference speaker(s)	-	SP_L		SP_L, SP_R		-	SP_L		SP_L, SP_R	
Apply the proposed method	-	no	yes	no	yes	-	no	yes	no	yes
Recognition rate (1 candidate) [%]	99.0	49.4	72.0	26.0	59.8	98.0	47.8	67.4	24.5	48.3
Recognition rate (5 candidates) [%]	100.0	67.9	88.1	42.4	77.0	99.8	66.8	83.4	41.7	70.5

recognition. The Cepstral Mean Normalization is applied when Julius extracts features of utterances.

Because the degradation of recognition rate affects the learners' motivation seriously [7], In addition to measure recognition rates with a candidate to each input, recognition rates with up to 5 recognition candidates to each input were also measured. Each recognition is considered to be succeed when these candidates include correct target word. This criteria is the same as the developed utterance training system.

D. Results

Fig. 8 and 9 show results of word recognition rate in case of anechoic room and students' room, respectively. Horizontal and vertical axes are word recognition rate and the number of interferences, respectively. Left bar indicates results before processing, and right one is those after processing.

All the results after processing show better performance than those before processing except for clean speech. In addition, the recognition rate is improved even though reverberant condition. Those results show that the proposed method is effective.

However, comparing those results with results by NTP [1],

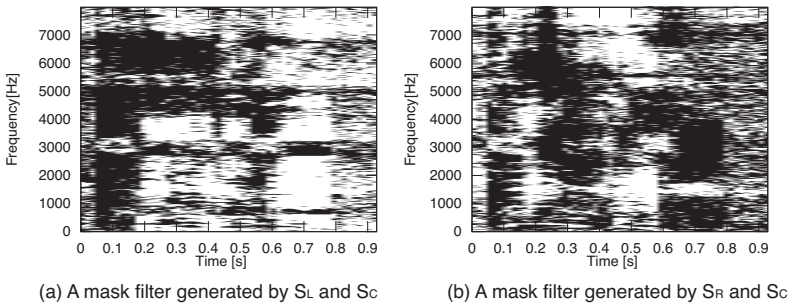

Fig. 5. Examples of mask filters generated by two signal pairs. (white: subtraction, black: addition)

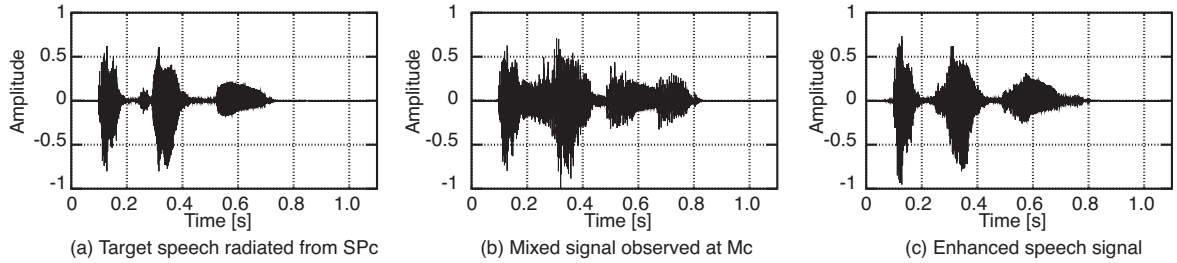

Fig. 6. An example of waveforms before and after processed. (/asahi/)

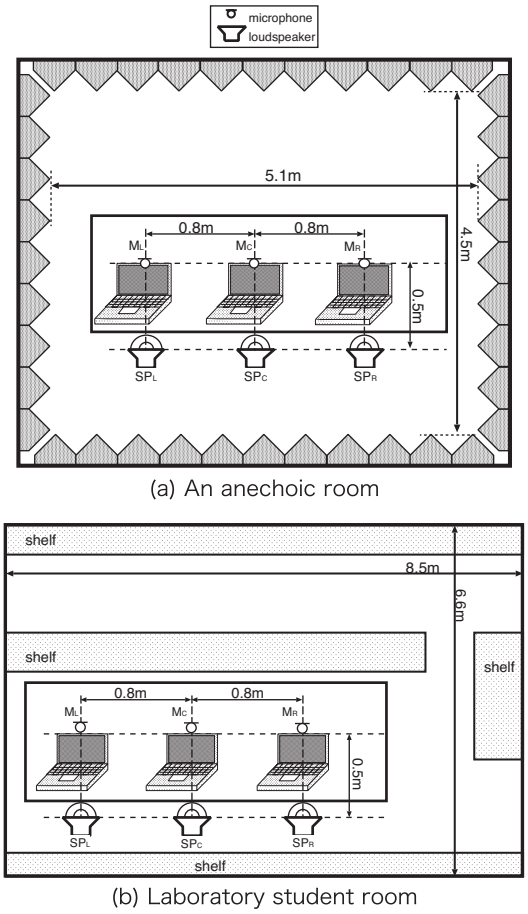

Fig. 7. Configurations of the simulation.

there are no apparent differences in recognition rate in both conditions.

V. Conclusions

In this paper, a speech enhancement method utilizing asynchronous multiple signals observed at microphones connected over TCP/IP network is proposed. PTP is introduced for a precise time synchronization on data acquisition.

Results of simulations show a possibility to improve the performance of speech recognition even though there are time lags in synchronization, reverberation and several interference speakers.

Comparing results by PTP with those by NTP, there were no significant differences nevertheless average error and standard deviation in time of PTP is better than those of NTP. Introducing of the specialized hardware is expected to improve performance on multi-channel signal processing, however, it costs so much and has difficulty to implement a precise clock module on a general portable device. Moreover, NTP does not require a specialized hardware and can be used in all the network. According to the results, robust multi-channel signal processing algorithm against a slight time lag should be discussed before making efforts to obtain a precise time synchronization.

Acknowledgment

Part of this work is carried out by Grant-in-Aid for Scientific Research (C) No. 24500147.

References

[1] T. Mashima, Y. Chisaki, T. Usagawa, "Speech Enhancement Method utilizing Asynchronous Signals with Time Code over TCP/IP Network," *Proc. Asia Pacific Signal and Information Processing Association Annual Summit a nd Conference 2010 Student Symposium,* p.13(CD-ROM) 2010.

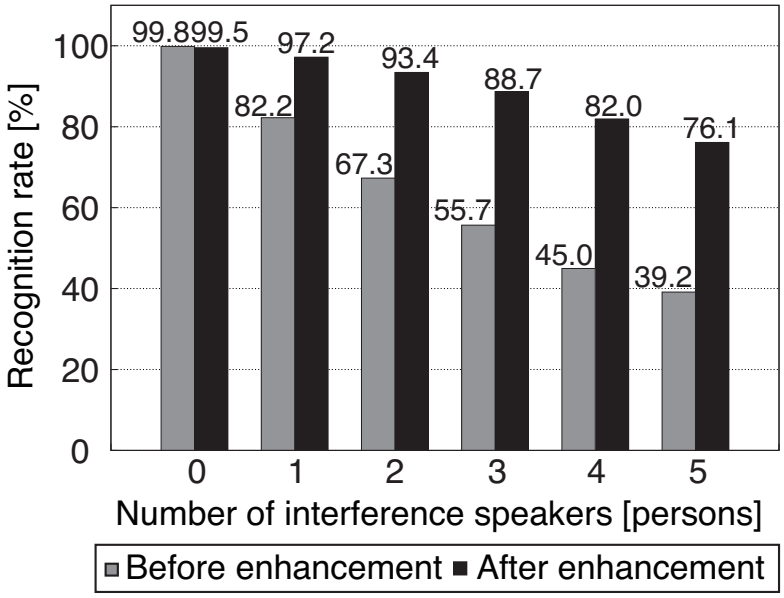

Fig. 8. Recognition rate against the number of interferences. (anechoic room, 5–best)

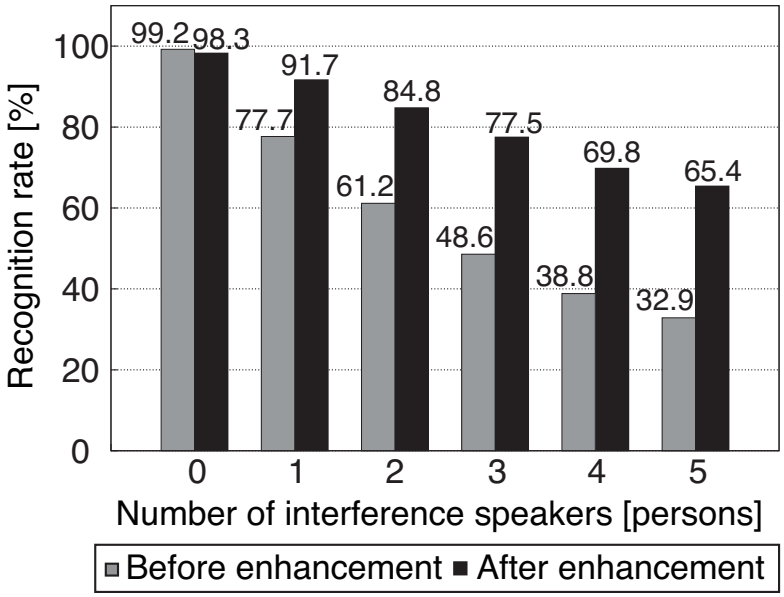

Fig. 9. Recognition rate against the number of interferences. (students' room, 5–best)

[2] G. Hu and D.L. Wang., Speech segregation based on pitch tracking and amplitude modulation, *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,* pp.79–82, Mohonk, NY, 2001.

[3] *ITU-T Recommendation P.340 "Transmission characteristics and speech quality parameters of hands-free terminals,"* May 2000.

[4] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses, *Journal of Acoustical Society of America*, Vol. 97, pp.1119–1123, 1995.

[5] S. Makino, K. Niyada, Y. Mafune and K. Kido,Tohoku University and Panasonic isolated spoken word database, *The Journal of the Acoustical Society of Japan 48(12),* pp.899–905, 1992.

[6] Lee, A., *Japanese large vocabulary speech recognition engine julius/julian*, http://julius.sourceforge.jp/

[7] J. Watanabe, K. Yamakawa, K. Umeda, Y. Chisaki and T. Usagawa, An Utterance Training System for L2 Learners of Japanese, Implemented on Moodle, *Proc. of the 8th International Conference on Information Technology Based Higher Education and Training, No. 91,* pp.1–5, 2007.