interspeech 2008

Development and Evaluation of Hands-Free Spoken Dialogue System for Railway Station Guidance

Hiroshi Saruwatari, Yu Takahashi, Hiroyuki Sakai, Shota Takeuchi, Tobias Cincarek, Hiromichi Kawanami, Kiyohiro Shikano

Nara Institute of Science and Technology, Nara, 630-0192 JAPAN

sawatari@is.naist.jp

Abstract

In this paper, we describe development and evaluation of hands-free spoken dialogue system which is used for railway station guidance. In the application at the railway station, noise robustness is the most essential issue for the dialogue system. To address the problem, we introduce two key techniques in our proposed hands-free system; (a) blind spatial subtraction array (BSSA) as a preprocessing, which can efficiently reduce nonstationary and diffuse noises in real-time, and (b) robust voice activity detection (VAD) based on speech decoding for further improvement of speech recognition accuracy. The experimental assessment of the proposed dialogue system reveals that the combination of real-time BSSA and robust VAD can provide the recognition accuracy of more than 80% under adverse railway-station noise conditions.

Index Terms: spoken dialogue system, hands-free, noise reduction, microphone array

1. Introduction

Spoken dialogue system is an essential technology for realizing an intuitive, unconstrained, and stress-free human-machine interface. Recently much attention has been paid in development of spoken dialogue system handled under real acoustical environments. As a good example, our spoken-oriented guidance system "Kitarobo" is working in an actual railway station since the end of March 2006 in Japan [1]. The system is located near the ticket gate, and everybody can use the system while the station is open. This system can provide guidance information to visitors regarding issues on the station itself and around the station, e.g., map and travel information (see [1] for details).

The input device of the original system was a close-talking microphone. Needless to say, this input style led to unnatural communication in that user must approach the microphone too closely, unlike human-human interface. To address this problem, in this paper, we extend our original one to be a hands-free spoken dialogue system. This improved system undertakes very challenging task, where there exist noises consisting of various kinds of interferences, e.g., background noise, sounds of trains, ticket-vending machines, automatic ticket wickets, foot steps, cars, and wind. They result in a *nonstationary and diffuse* noise environment, and thus it is too difficult to reduce the noises only using the conventional noise reduction methods such as single channel spectral subtraction (SS) [2] or simple beamforming technique via microphone array.

Two key techniques in our proposed hands-free system are (a) blind spatial subtraction array (BSSA) [3] as a preprocessing, which can efficiently deal with nonstationary and diffuse noises, and (b) robust voice activity detection (VAD) based on speech decoding [4]. Configuration of the whole dialogue system is shown in Fig. 1. In this paper, we newly propose a real-time architecture of BSSA and implement the real-time BSSA. Moreover, we introduce the implemented real-time BSSA and robust VAD method into the spoken-oriented guidance system. The experimental assessment of the proposed dialogue system reveals that the combination of real-time BSSA and robust VAD can provide the recognition accuracy of more than 80% under adverse railway-station noise conditions.

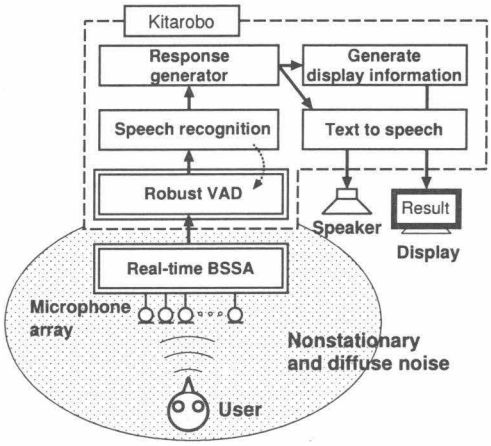

Figure 1: Overview of hands-free spoken dialogue system with real-time BSSA and robust VAD.

2. Blind spatial subtraction array

2.1. Motivation and feasibility

In the last decade, independent component analysis (ICA) [5] becomes one of the most notable candidate of microphone array method for separating and reducing interfering sounds in acoustical signal processing [6]. This is due to the feasible property that ICA is unsupervised adaptive signal processing, where training sequences, a priori information of the microphones' positions and their calibrations are not needed in advance. Generally speaking, the conventional ICA could work particularly in speech-speech mixing, i.e., all sound sources can be regarded as point sources. However, such a mixing assumption is very unrealistic in the railway-station environment, where the following scenario are likely to arise.

Target speech model: The target sound is user's speech, which can be approximately regarded as a *point source* locating relatively *close to the microphone array* (e.g., 1 or 2 m apart). Consequently the accompanying reflection and reverberation com-

Accepted after peer review of full paper

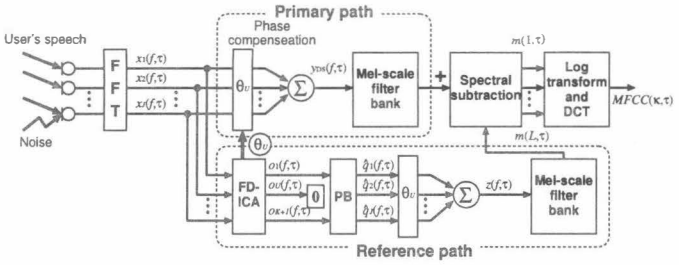

Figure 2: The block diagram of the off-line BSSA

ponents are small.

Noise model: We are often confronted with interference sounds which are *not point sources* but widespread sources. Also the noise is usually far from the array and heavily reverberant.

From the above-mentioned scenario, the conventional ICA can suppress the user's speech signal to pick up the noise source, but ICA is very weak in picking up target speech itself via suppression of the far-located widely-spread noise. This is because ICA with the small number of sensors and filter taps often provides only directional nulls against the undesired source signals [6]. This gives us an unfortunate conclusion that ICA is *not* proficient in speech enhancement in the railway-station noise environment. However, this also implies that we can still use ICA as an accurate noise estimator. Based on the above-mentioned fact, we decided to use our previously proposed BSSA [3] that utilizes ICA as a noise estimator. In BSSA, source extraction is achieved by subtracting the power spectrum of the estimated noise via ICA from the power spectrum of the target speech enhanced observed signal via delay-and-sum (DS). Although our previous work [3] only gives mathematical basis of BSSA algorithm and did not refer to diffuse noise reduction, BSSA's attractive feature is well expected to be suitable for an robust reduction of the railway-station noise.

2.2. Basic principle of BSSA

The block diagram of the BSSA is shown in Fig. 2. BSSA consists of two paths; a primary path which is DS-based target speech enhancer, and a reference path which is ICA-based noise estimator. Finally, we obtain the target speech extracted signal based on spectral subtraction procedure [2].

First, the observed signal vector in time-frequency domain is defined as

$$\boldsymbol{x}(f,\tau) = [x_1(f,\tau), \ldots, x_J(f,\tau)]^{\mathrm{T}}, \tag{1}$$

where $\boldsymbol{x}(f,\tau)$ is the observed signal vector, f is the frequency bin, τ $(= 0, 1, 2, \ldots)$ is time frame index, and J is the number of microphones. In the primary path, the target speech is partly enhanced via DS; the procedure can be given as

$$y_{\mathrm{DS}}(f,\tau) = \boldsymbol{g}_{\mathrm{DS}}(f,\theta_U)^{\mathrm{T}}\boldsymbol{x}(f,\tau), \tag{2}$$

$$\boldsymbol{g}_{\mathrm{DS}}(f,\theta) = [g_1^{(\mathrm{DS})}(f,\theta), \ldots, g_J^{(\mathrm{DS})}(f,\theta)]^{\mathrm{T}}, \tag{3}$$

$$g_j^{(\mathrm{DS})}(f,\theta) = \frac{1}{J}\exp\left(-i2\pi(f/M)f_{\mathrm{s}}d_j\sin\theta/c\right), \tag{4}$$

where $\boldsymbol{g}_{\mathrm{DS}}(f,\theta)$ is the coefficient vector of DS array, and θ_U is the look direction which is estimated by the unmixing matrix optimized by ICA [6]. Also, f_s is the sampling frequency and d_j $(j = 1, \cdots, J)$ is the microphone position. Besides, M is the DFT size, and c is the sound velocity.

In the reference path, the ICA-based noise estimation is performed. First, we perform signal separation using the complex valued unmixing matrix $\boldsymbol{W}_{\mathrm{ICA}}(f)$, so that the output signals $\boldsymbol{o}(f,\tau) = [o_1(f,\tau), \ldots, o_K(f,\tau)]^{\mathrm{T}}$ become mutually independent; this procedure can be represented by

$$\boldsymbol{o}(f,\tau) = \boldsymbol{W}_{\mathrm{ICA}}(f)\boldsymbol{x}(f,\tau), \tag{5}$$

$$\boldsymbol{W}_{\mathrm{ICA}}(f) = \begin{bmatrix} W_{11}^{(\mathrm{ICA})}(f) & \cdots & W_{1J}^{(\mathrm{ICA})}(f) \\ \vdots & \ddots & \vdots \\ W_{K1}^{(\mathrm{ICA})}(f) & \cdots & W_{KJ}^{(\mathrm{ICA})}(f) \end{bmatrix}. \tag{6}$$

Also, the unmixing matrix is updated iteratively by

$$\boldsymbol{W}_{\mathrm{ICA}}^{[p+1]}(f) = \mu\left[\boldsymbol{I} - E[\varphi(\boldsymbol{o}(f,\tau))\boldsymbol{o}^{\mathrm{H}}(f,\tau)]\right]\boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f) + \boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f), \tag{7}$$

where μ is the step size parameter, $[p]$ is used to express the value of the p-th step in the iterations, $\boldsymbol{I}$ is an identity matrix, and $E[\cdot]$ is the expectation operator. Besides, $\boldsymbol{M}^{\mathrm{H}}$ denotes hermitian transpose of matrix $\boldsymbol{M}$, and $\boldsymbol{\Phi}(\cdot)$ is the appropriate nonlinear vector function [6].

In the reference path, it is only required to estimate noise component. Thus, the target signal component $o_U(f,\tau)$ is removed from the output signal vector $\boldsymbol{o}(f,\tau)$. This processing can be designated as

$$\boldsymbol{q}(f,\tau) = [o_1(f,\tau), \ldots, o_{U-1}(f,\tau), 0, o_{U+1}(f,\tau), \ldots, o_K(f,\tau)]^{\mathrm{T}}. \tag{8}$$

Next, we apply the projection back (PB) method to remove the ambiguity of amplitude. This procedure can be represented as

$$\hat{\boldsymbol{q}}(f,\tau) = \boldsymbol{W}_{\mathrm{ICA}}^{+}(f)\boldsymbol{q}(f,\tau), \tag{9}$$

where $\boldsymbol{M}^{+}$ denotes Moore-Penrose pseudo inverse matrix of $\boldsymbol{M}$. Next, we obtain the estimated noise signal $z(f,\tau)$ by performing DS as follows:

$$z(f,\tau) = \boldsymbol{g}_{\mathrm{DS}}^{\mathrm{T}}(f)\hat{\boldsymbol{q}}(f,\tau). \tag{10}$$

Note that $z(f,\tau)$ is the function of the frame number τ, unlike the constant noise prototype estimated in the traditional SS [2]. Therefore, BSSA can deal with *nonstationary* noise.

Finally, source extraction is achieved by spectral subtraction as follows

$$y_{\mathrm{BSSA}}(f,\tau) = \begin{cases} |y_{\mathrm{DS}}(f,\tau)|^2 - \beta\cdot|z(f,\tau)|^2 \ ^{\frac{1}{2}}, \\ \quad (\text{ if } |y_{\mathrm{DS}}(f,\tau)|^2 - \beta\cdot|z(f,\tau)|^2 \geq 0\ ) \\ \gamma\cdot|y_{\mathrm{DS}}(f,\tau)| \quad (\text{otherwise}), \end{cases} \tag{11}$$

where $y_{\mathrm{BSSA}}(f,\tau)$ is the final output BSSA, β is the oversubtraction parameter, and γ is the flooring parameter. The appropriate setting, e.g., $\beta > 1$ and $\gamma \ll 1$, gives an efficient noise reduction.

2.3. Real-time implementation of BSSA

In BSSA's signal processing, DS, SS, and separation filtering parts are possible to work in real-time. However, it is toilsome to optimize (update) the separation filter in real-time because the optimization of the unmixing matrix by ICA consumes huge amount of computations. Therefore, in this paper we newly introduce a strategy in that the separation filter optimized by using the past time period data is applied to the current data. Figure 3 illustrates a configuration of the proposed real-time implementation of BSSA. Signal processing in this implementation is performed via the following manner.

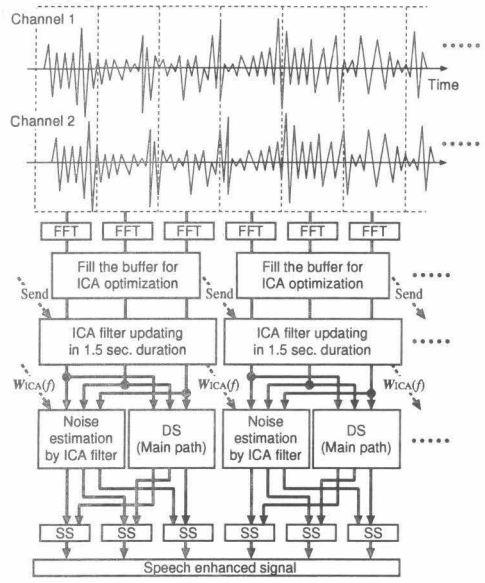

Figure 3: Signal flow in real-time implementation of BSSA.

Step 1: Inputted signals are converted into time-frequency domain series by using a frame-by-frame fast Fourier transform (FFT).

Step 2: ICA is conducted using the past 1.5-s-duration data for estimating separation filter while the current 1.5 s. The optimized separation filter is applied to the next (*not current*) 1.5 s samples. This staggered relation is due to the fact that the filter update in ICA requires substantial computational complexities and cannot provide the optimal separation filter for the current 1.5 s data.

Step 3: Inputted data is processed in two paths. In the primary path, target speech is partly enhanced by DS. In the reference path, ICA-based noise estimation is conducted. Again, note that the separation filter for ICA is optimized by using the past time period data.

Step 4: Finally, we obtain the target-speech-enhanced signal by subtracting the power spectrum of the estimated noise signal in the reference path from the power spectrum of the primary path's output.

Although the separation filter update in the ICA part is not real-time processing but involves totally a latency of 3.0 seconds, the entire system still seems to run in real-time because DS, SS and separation filtering can work in the current segment with no delay. In the system, the performance degradation due to the latency problem in ICA is mitigated by oversubtraction in the spectral subtraction.

3. Robust VAD based on speech decoding

Although BSSA can generally gain the output signal-to-noise ratio (SNR) of about 10∼15 dB, it is still insufficient to improve the speech recognition accuracy because SNR of the input speech is heavily degraded, e.g., up to 0 dB, in the actual railway-station noise environment. Consequently, the conventional VAD (based on amplitude level and zero cross counting) is likely to fail, and this will yield speech recognition failure.

To solve the low input-SNR problem, we have proposed a novel noise-robust VAD method based on speech decoding [4]. Our robust VAD method is embedded in a part of speech decoder different from the conventional VADs, namely, VAD and

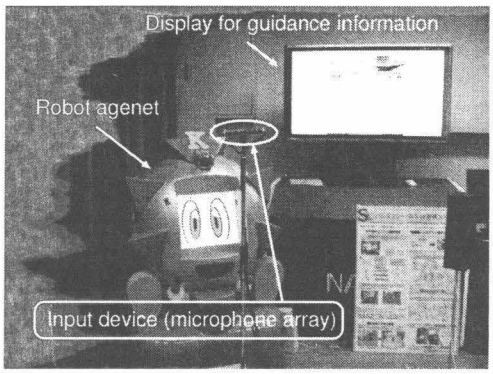

Figure 4: Appearance of hands-free spoken dialogue system.

speech decoding is performed in parallel. Robust VAD focuses on the premise that non-speech segments (silence) exist before and after the speech part of the utterance. Speech decoder can detects silence or noise segments. A frame or series of frames which are recognized as part of phonemes or series of phonemes is considered as voice activity. Otherwise, a sequence of silence frames of about 300–400 msec in duration is considered as noise. It has been reported that the method can increase the recognition performance in typical noisy environments (see [4] for more details). In this paper, we newly apply the robust VAD method into our development of hands-free spoken dialogue system in cooperation with real-time BSSA.

4. Experiment and performance assessment

4.1. Simulating railway-station noise

The main task of Kitarobo is a station guidance, and always working in an actual railway-station. Thus, it is difficult to conduct various assessment experiments in an arbitrary time. Therefore, we have a necessity to construct the noise environment simulator of railway-station for experiments. To solve the problem, we have constructed the experimental room for hands-free spoken dialogue system with the real-time BSSA. The experimental room contains Kitarobo with the real-time BSSA and railway-station noise simulator (see Fig. 4). The noise is simulated via recording noises in an actual railway station with eight-channel microphones, and playback of the multi-channel recorded railway-station noise by eight surrounded loudspeakers (see Fig. 5).

4.2. Experimental setup

To evaluate the hands-free spoken dialogue system with the real-time BSSA, the speech recognition test was conducted. Figure 5 depicts a layout of a reverberant room in our experiment where the reverberation time is more than 400 ms. The following real-recored 16 kHz-sampled signals were used in the experiments. The target signal is user's speech which is talked in front of a microphone array and 1.5 m apart from the array. As for noise, two noises were added simultaneously. First noise is the real-recored noise in an actual railway-station noise (it simulates railway-station noise) emitted from surrounded 8 loudspeakers. Second noise is an interference speech located at 50 degrees in the right direction of the microphone array, and its distance is 2.0 m.

We use 5 speakers (250 words) as target user, and Julius [7] ver. 4.0 RC2 as speech decoder. A eight-element array with the interelement spacing of 2 cm is used. The array consists

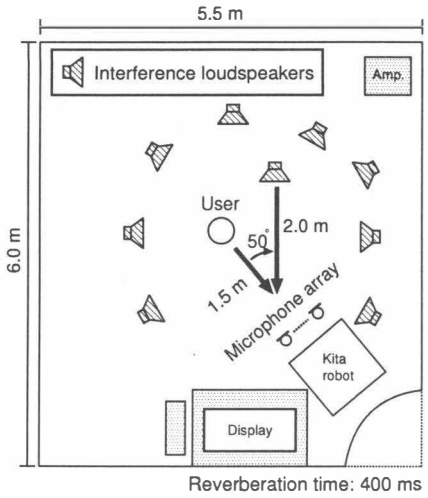

Figure 5: Layout of reverberant room in our experiment.

of directional microphone SHURE MX-184. DFT size is 512 points, window length for ICA is 256 points, and window shift size is 128 points in the experiment. Thus, the algorithm latency of the real-time BSSA is about 50 ms.

4.3. Experimental result and discussion

Figures 6 and 7 depict speech recognition results with or w/o interference speech, where we compare (a) conventional *single-mic.* system without robust VAD and BSSA, (b) the system using robust VAD only, and (c) the proposed system with robust VAD and BSSA. In the case without interfering speech, robust VAD mainly contributes to the improvement of recognition performance, but in the case including the interference speech, BSSA notably sustains the recognition accuracy of more than 80%. This is very natural because speech-decoding-based VAD depends on the existence of *unknown (untrained)* noise such as speech interference, but BSSA cannot be affected. Overall, the combination of BSSA and robust VAD is beneficial to hands-free spoken dialogue system under the adverse condition.

Figure 8 gives another comparative assessment from the viewpoint of microphone array signal processing. The results reveal that both the word correct and word accuracy of the proposed BSSA are obviously superior to those of DS and the conventional ICA; this is a promising evidence of the proposed signal processing's efficacy. The demonstration movie of our hands-free spoken dialogue system is available in the following URL. Readers can confirm that the fluent conversation including accurate responses with small latency is achieved via real-time BSSA and VAD.

Demo video: **http://spalab.naist.jp/database/Demo/rtbssa/**

5. Conclusion

In this paper, we propose the hands-free speech-oriented guidance system used in the railway-station noise environment. To handle the noise robustness, we introduce two key techniques, namely, real-time BSSA for efficient reduction of diffuse noises, and robust speech-decoding-based VAD for further improvement of recognition accuracy. The experimental results reveal that the combination of real-time BSSA and robust VAD can provide the recognition accuracy of more than 80% under adverse railway-station noise conditions. Also it is confirmed that real-time BSSA outperforms DS- and ICA-based methods.

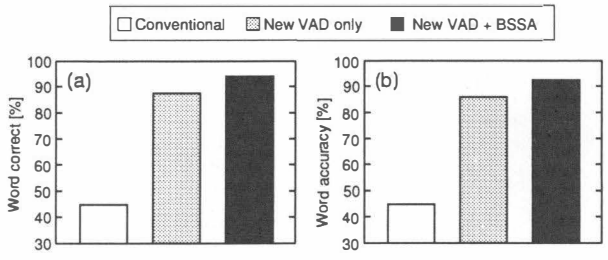

Figure 6: Contribution of VAD and BSSA in (a) word correct, and (b) word accuracy (without interference speech).

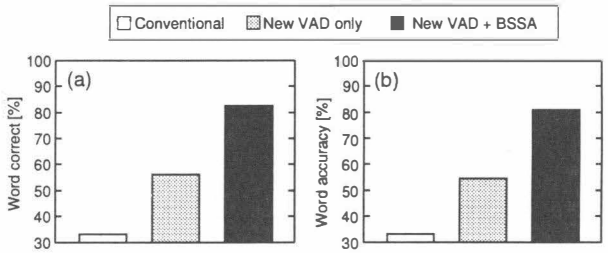

Figure 7: Contribution of VAD and BSSA in (a) word correct, and (b) word accuracy (with interference speech).

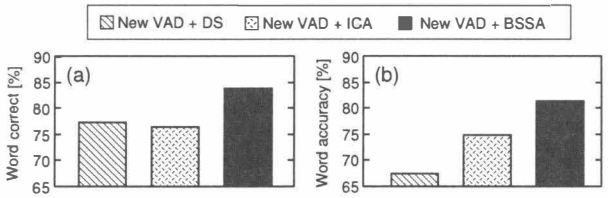

Figure 8: Comparison of signal processing methods in (a) word correct, and (b) word accuracy (with interference speech).

Acknowledgement: This work was partly supported by the NEDO project for strategic development of advance robotics elemental technologies in Japan.

6. References

[1] H. Kawanami, et al., "Development and operational result of real environment speech-oriented guidance systems kita-robo and kita-chan," *Oriental COCOSDA 2007*, pp.132–136, 2007.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, Signal Proc.*, vol.ASSP-27, no.2, pp.113–120, 1979.

[3] Y. Takahashi et al., "Blind spatial subtraction array with independent component analysis for hands-free speech recognition," *Proc. of IWAENC*, 2006.

[4] H. Sakai, et al., "Voice activity detection applied to hands-free spoken dialogue robot based on decoding using acoustic and language model," *ROBOCOMM2007*, 2007.

[5] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol.36, pp.287–314, 1994.

[6] H. Saruwatari, et al., "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Transactions on Speech and Audio Processing*, vol.14, no.2, pp.666–678, 2006.

[7] A. Lee, et al., "Julius – An open source real-time large vocabulary recognition engine," *Proc. EUROSPEECH*, pp.1691–1694, 2001.