HANDS-FREE SPEECH RECOGNITION CHALLENGE FOR REAL-WORLD SPEECH DIALOGUE SYSTEMS

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

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

ABSTRACT

In this paper, we describe and review our recent development of hands-free speech dialogue system which is used for railway station guidance. In the application at the real railway station, robustness against reverberation and noise 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) speech dialogue system construction with real speech database collection and language/acoustic model improvement, and (b) microphone array preprocessing using blind spatial subtraction array which can solve the reverberation-naiveness problem inherent in conventional microphone arrays. The experimental assessment of the proposed dialogue system reveals that our system can provide the recognition accuracy of more than 80% under realistic railway-station conditions.

Index Terms— Speech dialogue, hands-free, real environment, microphone array, reverberation robustness

1. INTRODUCTION

Speech 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 speech dialogue systems handled under real acoustical environments. Needless to say, close-talking input style, which was commonly used in most of the speech recognition researches, led to unnatural communication in that user must approach the microphone too closely, unlike human-human interface. Hence *hands-free* speech recognition/dialogue systems are now being studied to undertake very challenging task, where there exist reverberation and interfering noises which heavily deteriorate target speech quality.

In this paper, we describe and review an issue of our recently developed hands-free speech dialogue systems. Particularly we focus on solutions for addressing the real-world acoustic problems due to the reverberation and noise. Key techniques in our proposed hands-free system are summarized in the following points: (a) Speech dialogue system construction [1, 2] operated in real environments, including real speech database collection and language/acoustic model improvement. (b) Microphone-array-based preprocessing structure using blind spatial subtraction array (BSSA) [3], which can effectively deal with reverberation-naiveness problem inherent in conventional microphone arrays.

Following this section, first, we mention a detailed review on speech dialogue system construction in Sect. 2. Next, in Sect. 3, we give a new analytical report on a BSSA's ability and advantage in the noise estimation from the viewpoint of reverberation robustness; this analysis provides a valuable insight for all studies on microphone array signal processing used in real world. Finally, we introduce the hands-free railway-station guidance system consisting of BSSA and the speech dialogue system. The experimental assessment of the proposed dialogue system reveals that our system can provide the recognition accuracy of more than 80% under realistic railway-station conditions.

This work was partly supported by the NEDO project for strategic development of advance robotics elemental technologies, and MIC Strategic Information and Communications R&D Promotion Programme in Japan.

2. REAL ENVIRONMENT SPEECH DIALOGUE SYSTEMS

2.1. Background

Since Nov. 2002, the authors have been operating a real environment speech-oriented guidance system named *Takemaru-kun* at an entrance of a community center. Following the success of *Takemaru-kun*, two guidance systems *Kita-chan* and *Kita-robo* are developed and implemented at a railway station, which is more noisy and reverberant environment than that of *Takemaru-kun*. In this section, we describe an overview of the systems, the speech database and acoustic models we have constructed for real environment operation.

2.2. System overview

Takemaru-kun is installed at an entrance lobby of the public center in Ikoma city, Nara [1]. *Kita*-systems (*Kita-chan* and *Kita-robo*) are set at *Gakken-kita-ikoma* railway station after Mar. 2006 [2]. To realize a natural human-machine communication, a robot-like body is equipped with *Kita-robo*. Figures 1 and 2 illustrate the snapshots of *Takemaru-kun* and *Kita-robo*. Noise levels of these locations are about 50 dBA and 60 dBA in the community center and the railway station, respectively. A directional microphone is used for these real environment systems to pick up users' utterances clearly.

In *Takemaru-kun* and *Kita*-systems, the system structure and dialogue strategy are fundamentally the same but service contents differ because of difference of locations and interfaces. For example, *Takemaru-kun* replies information about the facilities, sightseeing, weather forecast, and so on. The other two systems can deal with transfer information in addition to the above.

2.3. System structure

The systems employ call-routing type response generation and example-based one-question-one answer strategy. A system output is given via synthetic speech as well as Internet web browser and CG animation. The process flow of speech recognition and response generation is illustrated in Fig. 3.

A system input is classified into speech or noise category based on five Gaussian mixture models (GMMs); adult speech, child speech, laugh, cough and noise. If an input is not classified as speech, the input is rejected [4].

Adult and child discrimination is also realized by comparing the acoustic likelihoods [5]. The discrimination result is used to select an appropriate answer concerning age groups. In the systems, four

Fig. 1. Appearance of *Takemaru-kun*.

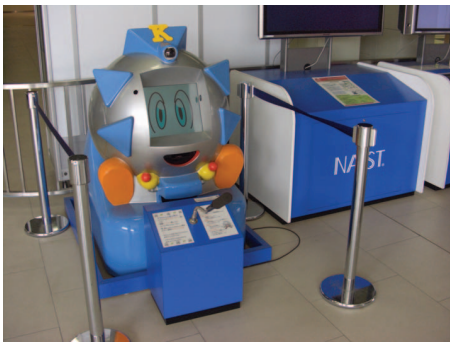

Fig. 2. Appearance of *Kita-robo*.

speech decoding processes run in parallel combining acoustic models (adult or child) and language models (network grammars and statistical N-gram models). Candidates with high acoustic likelihood are selected as a final recognition result.

A system response sentence is generated using a question and answer database (QADB) for N-gram decoding and prepared rules for network grammar decoding. In the example-based approach, an example question which is most similar to the system input is selected from QADB.

The corresponding response in QADB is outputted as the system answer. Two QADBs are installed for responding adult and child separately. Although the initial QADB for *Takemaru-kun* is given manually, they have been updated by adding users' real speech to the systems. To construct a precise QADB, an automatic optimization method has been reported in [6].

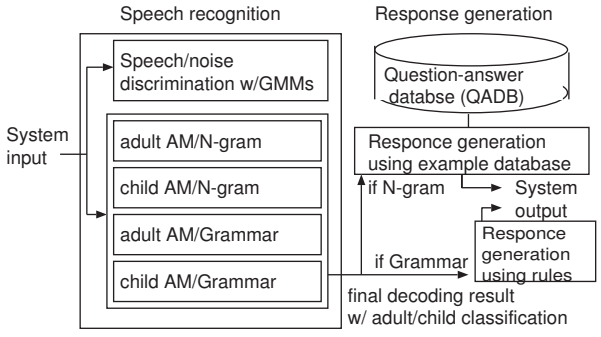

Fig. 3. Basic structure of dialogue systems.

Table 1. Speech data collected with *Takemaru-kun* and *Kita*-systems by the end of Dec. 2007

Classification	*Takemaru-kun* # Inputs	*Takemaru-kun* Time	*Kita*-systems # Inputs	*Kita*-systems Time
Transcribed	273,698	121.2 h	82,845	41.1 h
Preschool children	27,537	14.3 h	10,115	5.4 h
Lower grade children	106,797	57.7 h	25,563	14.2 h
Higher grade children	31,402	15.8 h	9,980	4.8 h
Adults, elderly	31,100	14.1 h	24,835	10.8 h
Noise, non-verbals	76,864	19.3 h	12,352	5.1 h
Untranscribed	684,461	334.6h	186,830	107.6 h
Total	958,159	455.8 h	269,675	148.7 h

2.4. Real-environment speech database

All inputs to the systems have been recorded since their initial operation. The first two-year speech data of *Takemaru-kun* and the first 10-month speech data of *Kita*-systems are manually transcribed with labels of age group, gender and tags about noise. Statistics of the number of inputs collected by end of December 2007 with *Takemaru-kun* and *Kita*-systems are shown in Table 1. When taking all systems together, more than 1.2 million inputs or more than 600 hours of real-environment speech and noise data have been collected. There are more than 270,000 speech and noise inputs or 120 hours of human-transcribed data from *Takemaru-kun* and more than 80,000 inputs or 40 hours from *Kita*-systems. These speech database are used to improve acoustic models, language models and QADBs. The quantitative analysis on the relationship among system ability, data size, and portability of the database in different dialogue systems are indicated in [7].

2.5. Acoustical model for handling real-world operation

In order to develop a real-environment system, acoustic model adaptation is conducted using railway-station background noise and room impulse responses.

As an initial acoustic model, conventional models in *Takemaru* are utilized. The models are already adapted to a dialogue system using 2-year data from *Takemaru-kun*. To employ the models to real-operation systems, railway-station background noise and impulse response are recorded. Six-month data from *Kita*-systems with impulse response convolution and noise addition are used for the adaptation based on MLLR-MAP (3 iterations).

3. REVERBERATION-ROBUST MICROPHONE ARRAY

3.1. Reverberation naiveness in conventional beamforming

In this section, we discuss a reverberation-related problem inherent in a conventional microphone array framework, especially on adaptive beamforming. The conventional adaptive beamformer, e.g., generalized sidelobe canceler [8], usually requires accurate noise estimation using a blocking spatial filter, i.e., *null beamformer* (NBF) [9]. The NBF-based noise estimator, however, suffers from the adverse effect of the room reverberation. NBF is a technique to suppress an objective source signal by generating a sharp null against the direction of the source signal. If the target speech signal arrives from the same direction as the steered null, we can suppress the

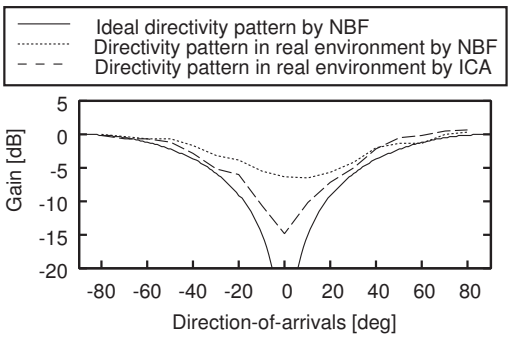

Fig. 4. Directivity patterns shaped by NBF and ICA in ideal and real environment where the reverberation time is 200 ms.

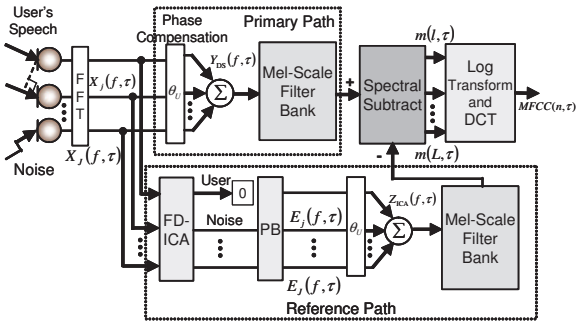

Fig. 5. Block diagram of BSSA.

speech signal perfectly and then we can estimate noise-only components. Due to the reverberation, however, the speech signal arrives from not only the null's direction but also outside of the direction. Therefore, in the reverberant room, we cannot suppress the speech signal sufficiently.

Figure 4 illustrates examples of *real directivity patterns* which are shaped by two-element NBF in the ideal (solid line) and the real (dotted line) environment where the reverberation time (RT) is 200 ms. These directivity patterns depict spacial gain responses for multiple directions, which are calculated by using real-recorded impulse responses in multiple directions. Thus they are *reverberation-included* responses. In this figure, the null direction is set to zero degree (normal to the array). We can see that the depth of the null in the real environment shallows because of the reverberation effect. Therefore, we cannot suppress the speech signal completely in the real environment by using NBF. In fact, NBF only provides a bad estimate of the noise signal with many speech-component leakages, and this results in *underestimate* of the speech signal in the following noise reduction step. Thus an improvement of reverberation-robustness in the noise estimator part is a problem demanding prompt attention.

3.2. Toward handling reverberation in microphone array

We have proposed an improved microphone array structure, blind spatial subtraction array (BSSA) [3], which includes *independent component analysis (ICA)-based noise estimator* instead of NBF-based noise estimator to address the reverberation-naiveness problem.

The block diagram of BSSA is shown in Fig. 5. BSSA consists of two paths; a primary path which is delay-and-sum (DS)-based target speech enhancer, and a reference path which is the ICA-based noise estimator. Finally, we obtain the target speech extracted signal based on spectral subtraction procedure. As for the reference path, we newly introduce ICA as a reverberation-robust noise estimator for adapting the spatial null filter to the reverberation (see Fig. 5).

In ICA, an unmixing matrix is optimized so that output signals become mutually independent only using observed signals, and a priori information about the direction of speech and the room acoustics is not required. Therefore the proposed method can estimate noise signals correctly regardless of the reverberation existence. Detailed signal processing is shown below.

In ICA part, we perform signal separation using the complex valued unmixing matrix $\boldsymbol{W}_{\mathrm{ICA}}(f)$, so that the output signals $\boldsymbol{O}(f,\tau) = [O_1(f,\tau),\ldots,O_J(f,\tau)]^{\mathrm{T}}$ become mutually independent; this procedure can be represented by

$$\boldsymbol{O}(f,\tau) \quad = \quad \boldsymbol{W}(f)\boldsymbol{X}(f,\tau) = \boldsymbol{P}(f)\boldsymbol{W}_{\mathrm{ICA}}(f)\boldsymbol{X}(f,\tau), \qquad (1)$$

where $\boldsymbol{P}(f)$ is a permutation matrix and $\boldsymbol{W}(f)$ is a new unmixing matrix in which the permutation problem is resolved. The permutation matrix $\boldsymbol{P}(f)$ is determined by looking at null directions in the directivity pattern which is shaped by $\boldsymbol{W}_{\mathrm{ICA}}(f)$ [9], so that the U-th output $O_U(f,\tau)$ is set to the target speech signal. The optimal $\boldsymbol{W}_{\mathrm{ICA}}(f)$ is obtained by the following iterative updating equation [10]:

$$\boldsymbol{W}_{\mathrm{ICA}}^{[p+1]}(f) = \mu\left[\boldsymbol{I} - \langle \boldsymbol{\Phi}(\boldsymbol{O}(f,\tau))\,\boldsymbol{O}^{\mathrm{H}}(f,\tau)\rangle_\tau\right]\boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f) + \boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f), \qquad (2)$$

where μ is the step-size parameter, $[p]$ is used to express the value of the p-th step in the iterations, and $\boldsymbol{I}$ is an identity matrix. Besides, $\langle\cdot\rangle_\tau$ denotes a time-averaging operator, $\boldsymbol{M}^{\mathrm{H}}$ denotes conjugate transpose of matrix $\boldsymbol{M}$, and $\boldsymbol{\Phi}(\cdot)$ is the appropriate nonlinear vector function [9]. In the reference path, the target speech component is discarded because we want to estimate only the noise component. Accordingly we remove the separated speech component $O_U(f,\tau)$ from ICA outputs $\boldsymbol{O}(f,\tau)$, and construct the following *noise-only vector* $\boldsymbol{Q}(f,\tau)$;

$$\boldsymbol{Q}(f,\tau) = [O_1(f,\tau),\ldots,O_{U-1}(f,\tau),0,O_{U+1}(f,\tau),\ldots,O_J(f,\tau)]^{\mathrm{T}}. \qquad (3)$$

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

$$\boldsymbol{E}(f,\tau) \quad = \quad \boldsymbol{W}^{+}(f)\boldsymbol{Q}(f,\tau). \qquad (4)$$

Here, $\boldsymbol{Q}(f,\tau)$ is composed of only noise components. Therefore, $\boldsymbol{E}(f,\tau)$ is a good estimation of the received noise signals at the microphone positions. Finally, we obtain the estimated noise signal $Z_{\mathrm{ICA}}(f,\tau)$ by performing DS as follows:

$$Z_{\mathrm{ICA}}(f,\tau) \quad = \quad \boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{E}(f,\tau), \qquad (5)$$

where $\boldsymbol{W}_{\mathrm{DS}}(f)$ is the weight vector of DS. Equation (5) is well expected to be equal to the noise term of the primary path.

3.3. Example of noise estimation experiment

Here we show one example of noise estimation performed under a real reverberant environment (RT = 200 ms) where target speech is contaminated by a cleaner noise. The broken line in Fig. 4 depicts the directivity pattern of ICA in the real environment. From this result, we can confirm that the null shaped by ICA becomes deeper compared with that of the NBF-based noise estimator. Next, Fig. 6 shows the long-term-averaged power spectra of the estimated noise

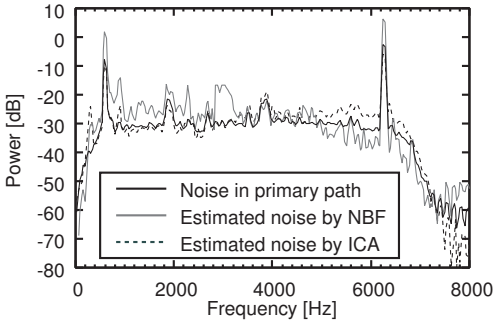

Fig. 6. Accuracy of estimated noise spectra by NBF and ICA.

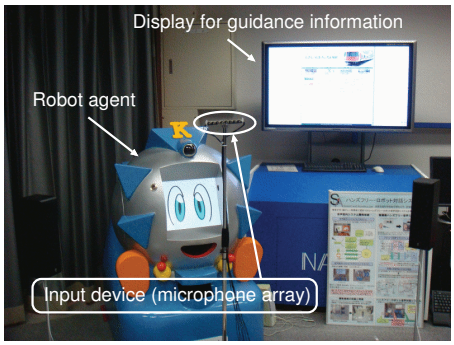

Fig. 7. Appearance of hands-free spoken dialogue system.

signals by NBF and ICA. We can see that the power spectrum of the estimated noise signal by NBF is not accurate but that of ICA is a good estimation, and hence reverberation naiveness is mitigated by ICA effectively.

4. DEVELOPMENT OF HANDS-FREE SPEECH DIALOGUE SYSTEM

Recently, we developed a hands-free speech dialogue system combining our original *kita*-systems and real-time BSSA [11], which is mainly aimed to be used for railway-station guidance in reverberant and noisy environment (see Fig. 7).

To evaluate the hands-free speech dialogue system, the speech recognition test was conducted in the reverberant room where the reverberation time is more than 400 ms. The target signal is user's speech which is talked in front of a microphone array and 1.5 m apart from the array. We use 5 speakers (250 words) as the target utterances. As for noise, two noises were added simultaneously. First noise is a diffuse noise recorded in an actual railway station emitted from surrounded 8 loudspeakers (it simulates railway-station noise). 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. An eight-element array with the interelement spacing of 2 cm is used.

Figure 8 gives a comparative assessment example from the viewpoint of preprocessing microphone array methods, i.e., we compare several candidates of hands-free speech dialogue system, *Kita-robo* with the conventional DS, ICA, or the proposed BSSA. The results reveal that both the word correct and word accuracy of the proposed BSSA are obviously superior to those of the conventional DS and ICA, and our proposed system notably sustains the recognition accuracy of more than 80%. This is a promising evidence of the proposed system'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 is achieved under the reverberant condition.

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

5. CONCLUSION

In this paper, we described the hands-free speech-oriented guidance system used in the railway-station noise environment. To handle the reverberation/noise robustness, we introduce two key techniques, namely, real-world speech dialogue system and reverberation-robust BSSA. The experimental results reveal that the combination of real-time BSSA and *kita*-system can provide the recognition accuracy of more than 80% under adverse railway-station noise conditions.

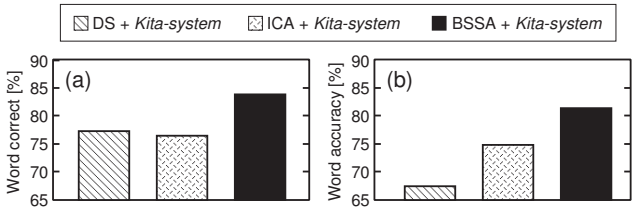

Fig. 8. Comparison of preprocessing methods in (a) word correct, and (b) word accuracy.

6. REFERENCES

[1] R. Nisimura, et al., "Operating a public spoken guidance system in real environment", *Proc. INTERSPEECH*, pp.845–848, 2005.

[2] H. Kawanami, et al., "Development and operational result of real environment speech-oriented guidance systems Kita-robo and Kita-chan," *Proc. oriental COCOSDA meeting*, pp.132–136, 2007.

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

[4] A. Lee, et al., "Noise robust real world spoken dialogue system using GMM based rejection of unintended inputs", *Proc. INTERSPEECH*, pp.173–176, 2004.

[5] R. Nisimura, et al., "Public speech-oriented guidance system with adult and child discrimination capability", *Proc. ICASSP*, pp.433–436, 2004.

[6] S. Takeuchi, et al., "Question and answer database optimization using speech recognition results," *Proc. INTERSPEECH*, pp.451–454, 2008.

[7] T. Cincarek, et al., "Development, long-term operation and portability of a real-environment speech-oriented guidance systems," *IEICE Trans. vol.E91-D, no.3,* pp.576–587, 2008.

[8] O. Hoshuyama, et al., "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Process.*, vol.47 no.10, pp.2677–2684, 1999.

[9] H. Saruwatari, et al., "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. Speech & Audio Process.*, vol.14, pp.666–678, 2006.

[10] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol.22, pp.21–34, 1998.

[11] Y. Takahashi, et al., "Real-time implementation of blind spatial subtraction array for hands-free robot spoken dialogue system," *Proc. IROS*, pp.1687–1692, 2008.