Robust Speech Recognition System for Communication Robots in Real Environments

Carlos Toshinori Ishi[†], Shigeki Matsuda[††], Takayuki Kanda[†], Takatoshi Jitsuhiro[†††], Hiroshi Ishiguro[†], Satoshi Nakamura[††], and Norihiro Hagita[†]

[†] Intelligent Robotics and Communication Laboratories, ATR, Kyoto, Japan

[††] National Institute of Information and Communications Technology, Japan; Spoken Language Communication Research Laboratories, ATR, Kyoto, Japan

[†††] Knowledge Science Laboratories, ATR, Kyoto, Japan

{carlos, shigeki.matsuda, kanda, takatoshi.jitsuhiro, satoshi.nakamura, hagita}@atr.jp, ishiguro@ams.eng.osaka-u.ac.jp

***Abstract* – The application range of communication robots could be widely expanded by the use of an automatic speech recognition (ASR) system with improved robustness for noise and for speakers of different ages. In this paper, we describe an ASR system which can robustly recognize speech by adults and children in noisy environments. We evaluate the ASR system in a communication robot placed in a real noisy environment. Speech is captured using a twelve-element microphone array arranged in the robot chest. To suppress interference and noise and to attenuate reverberation, we implemented a multi-channel system consisting of an outlier-robust generalized sidelobe canceller (RGSC) technique and a feature-space noise suppression using MMSE criteria. Speech activity periods are detected using GMM-based end-point detection (GMM-EPD). Our ASR system has two decoders for adults' and children's speech. The final hypothesis is selected based on posterior probability. We then assign a generalized word posterior probability (GWPP)-based confidence measure to this hypothesis, and if it is higher than a threshold, we transfer it to a subsequent dialog processing module. The performance of each step was evaluated for adults' and children's speech, by adding different levels of real environment noise recorded in a cafeteria. Experimental results indicated that our ASR system could achieve over 80 % word accuracy in 70 dBA noise. Further evaluation of adult speech recorded in a real noisy environment resulted in 73 % word accuracy.**

Index Terms – Communication robots, speech recognition, robustness, acoustic noise, children speech.

I. INTRODUCTION

Our research aims to develop "communication robots" that can naturally interact with humans and support everyday activities. Since the target audience of a communication robot is the general public who does not have specialized computing and engineering knowledge, a conversational interface using both verbal and non-verbal expressions is becoming more important. Previous studies in robotics have emphasized the merit of robot embodiments, showing the effectiveness of facial expression [1], eye-gaze [2], and gestures [3].

Recently, several practical robots have been developed, such as therapy tools [4], museum orientation tool [5], and entertainments [6]. Moreover, robots are enlarging their working field in our daily lives. In one of our previous work, we tested a child-size interactive humanoid robot at an elementary school for several weeks [7]. The robot interacted with children by using speech and gestures in a free play situation. In one of the interactions, the robot motivated children to learn English by talking in English to them [7]. A similar project was run in a science museum where a humanoid robot interacted with visitors in a free-play situation and also conducted a museum tour, which contributed to visitors to grow interests in science and technologies [8].

One criticism to these two field trials was that these robots lacked speech recognition capability. The robots interacted with humans by speaking and making gestures, which are important elements for creating a sense of reality in humanoid robots. Language-based communication is indispensable, in order to fully utilize their human-like presence. However, one of the difficulties concerned speech recognition in noisy environments. Current technology has a good performance in recognizing formal utterances in noiseless environments, but the performance drastically degrades in noisy environments.

Several researchers are recently endeavoring to solve such problem so called "robot audition" [9]-[13]. Most of these works makes use of microphone array technology, for realizing sound source localization and separation, prior to speech recognition. However, the evaluation is usually done by controlling the direction of the noise or the interference.

Further, although most works evaluating speech recognition by robots have focused only on adult speech, these field trials (in both elementary school and science museum), indicated that children are important robot users, as well as adults. Thus, such communication robots should be able to deal with speech recognition of both adults' and children's speech. However, the performance of speech recognition also degrades due to differences on speaker age.

In this paper, we evaluate our ASR (automatic speech recognition) system, which accounts for these two problems (caused by noisy environments and differences on speaker age), in a real noisy environment. Although we are conscious that a full communication could be reached by considering both verbal and non-verbal communication [14], in this paper,

we focus only on the evaluation of the verbal information processing.

The rest of the paper is organized as follows. In Section II, we introduce our ASR system, and describe the techniques used in each module. In Section III, we present the recognition performance results for several system structures, and for several noise conditions. We offer our conclusions in Section IV.

II. System Description

Accounting for the two problems (caused by noisy environments and differences on speaker age) described in Section I, we developed an ASR system to be robust to both background noise and speakers of different ages. Fig. 1 shows the structure of our ASR system. It consists of two major blocks.

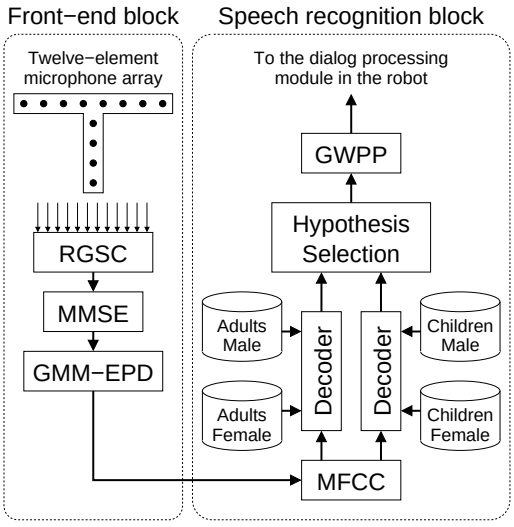

Fig. 1 The structure of the ASR system robust to noise and speakers of different age.

The first block is a front-end processing. It contains a twelve-element microphone array, as depicted in Fig. 2. The real-time multichannel system for suppressing interference and noise and for attenuating reverberation consists of an outlier-robust generalized sidelobe canceller (RGSC) and a feature-space noise suppression (MMSE). MMSE noise suppression is applied after RGSC to reduce the residual noise at the RGSC output. After that, the speech activity period detected by the GMM-based end-point detection (GMM-EPD) is transferred to the second block.

In the second block, there are two decoders depending on the age of the speaker (adult or child); each decoder works using gender-dependent acoustic models. Noise-suppressed speech at the first block is recognized using these two decoders, and one hypothesis is selected based on posterior probability. Finally, the hypothesis is measured using a generalized word posterior probability (GWPP)-based confidence measure. The hypothesis with confidence score higher than a threshold can then be transferred to a subsequent dialog processing module. The following sub-sections describe each module of our ASR system.

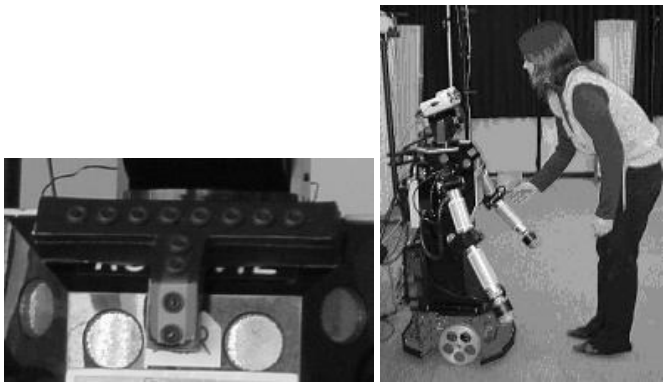

Fig. 2 Twelve-element microphone array in the Robovie's chest. Robovie wears the microphone array on its chest.

A. *Twelve-element microphone array*

In our system, we use a twelve-element microphone array for capturing speech. Omni-directional condenser microphones of type DPA 4060 are used for high-quality sound capture of distant-talking speech. The microphones are arranged in a T-shape with eight microphones on the horizontal axis and four microphones along vertical axis, with a spacing of 2-cm, as shown in Fig. 2.

We decided to arrange the microphone array in the robot chest, instead of the ear position or the head of the robot, for two reasons. One is the geometric limitations of the robot head, which would constraint the effective frequency range of the array processing. The other reason is that our robot makes rapid head movements, which would make difficult to set a target direction for the array processing.

B. *Outlier-robust generalized sidelobe canceller (RGSC)*

Many sound source separation algorithms have been proposed in order to reduce background noise coming from different directions. Here, we use an outlier-robust generalized sidelobe canceller (RGSC), proposed in [15]. The RGSC is applied to the audio signals captured using the twelve-element microphone array. The RGSC system is composed by a fixed beamformer, an adaptive blocking matrix, and an adaptive interference canceller, as shown in Fig. 3.

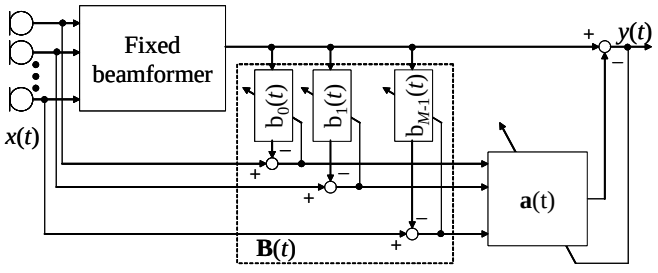

Fig. 3 Block diagram of RGSC.

The fixed beamformer steers the sensor array to the direction of the desired source and enhances the desired signal relative to the surrounding interference and noise. A simple uniformly weighted delay & sum beamformer is used. The fixed beamformer forms the reference path of the GSC. The blocking matrix **B**(t) is an adaptive spatial filter which suppresses the desired signal and which passes interference and noise, such that the output of **B**(t) is a reference for interference and noise. The adaptive interference canceller **a**(t) is realized by a multichannel adaptive filter between the output of the blocking matrix and the output of the fixed beamformer. The estimative of interference and noise is subtracted from the reference path at the output of the fixed beamformer so that the suppression of interference and noise is maximized.

The blocking matrix should be adapted when the signal-to-noise ratio (SNR) is high, while the interference canceller should be adapted when the SNR is low to prevent instability of the adaptive filters. A DFT bin-wise classifier for 'desired signal only', 'interference only' and 'double-talk' between the desired signal and interference or noise is then used for optimally tracking the desired signal and the interference. An "outlier-robust" adaptive filtering in the DFT domain for bin-wise adaptation control derived from [16] is then used for maximize the robustness against errors in the DFT bin-wise classifier.

C. Feature-space noise suppression using clean speech GMMs

The feature-space single-channel noise suppression, shown in Fig. 4, is applied after the RGSC to reduce the residual noise at the RGSC output.

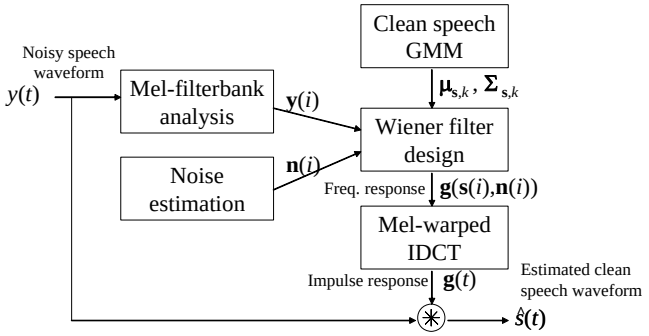

Fig. 4 Block diagram of the feature-space noise suppression.

A GMM (Gaussian Mixture Model)-based MMSE (Minimum Mean Square Error) estimator is used to estimate a Wiener filter for suppressing background noise. The feature space is constituted by log Mel-spectral energy coefficients. For each frame *i*, one Wiener filter is obtained as a linear interpolation of multiple sub-Wiener filters, which are calculated using individual mixture component *k* of a clean speech GMM ($\boldsymbol{\mu}_{s,k}$, $\boldsymbol{\Sigma}_{s,k}$), and the observed noise $\mathbf{n}(i)$. The weights of the multiple sub-Wiener filters are optimized, based on the MMSE criteria, by maximizing the likelihood between te clean speech GMM and the input speech noise-suppressed by the Wiener filter. The filtered signal $\mathbf{g}(\mathbf{s}(i),\mathbf{n}(i))$ obtained in the log Mel-frequency domain is transformed back to the time domain for obtaining the impulse response $g(t)$. The clean speech is then estimated by convoluting the input noisy speech $y(t)$ with the impulse response $g(t)$. More details about the evaluation of the present noise suppression module can be found in [15].

D. GMM-Based End-Point Detection

An End-Point Detection (EPD) module is necessary for communication robots to properly interact with the user. In our ASR system, a GMM-based end-point detection (GMM-EPD) is used for detecting speech activity periods. This type of EPD architecture is widely used as a noise-robust EPD. First, we estimate the GMMs of noisy speech and noise in advance using a sufficient amount of training data.

During the detection of speech activity periods, we calculate the likelihood between each GMM and the input noisy speech. If the likelihood of the noisy speech GMM is higher than the noise GMM, the current frame is labeled as speech.

E. Hypothesis selection

For the speech recognition decoder engine, we use a hypothesis selection technique based on posterior probability [17] for improving robustness to speakers of different ages. One advantage of such approach is that it does not need to previously recognize the speaker age. Instead, the hypothesis with the highest score is selected from multiple hypotheses as follows:

$$\hat{k} = \arg\max_{k=1}^{K} H_k \,, \tag{1}$$

$$H = \log P(\boldsymbol{X}|\boldsymbol{W}) + \lambda \log P(\boldsymbol{W}), \tag{2}$$

where H_k is the score of the hypothesis obtained from the *k*-th decoder and *K* denotes the number of decoders. The hypothesis obtained from the $\hat{k}$-th decoder has the highest score, which is defined as the sum of the log acoustic model likelihood $\log P(\boldsymbol{X}|\boldsymbol{W})$ and the log language model probability $\log P(\boldsymbol{W})$ of a hypothesis. $\boldsymbol{X}$, $\boldsymbol{W}$, and H are the observed feature vector sequence, the hypothesis represented by a word sequence, and the score for the hypothesis, respectively. λ denotes a language model weight used for the hypothesis selection. Details about evaluation of the speech recognition decoder module can be found in [17].

F. GWPP-based word confidence and rejection

So far, several techniques were described for improving the robustness of the ASR system to noise and to speakers of different ages. Nevertheless, the performance of an ASR system may degrade due to a mismatch between the training and testing channels, interference from environmental noise, etc. If the recognition results contain some fatal errors, this will adversely affect or prevent natural interaction between the

robot and a human. To measure the reliability of the recognition results of the ASR system, we use a generalized word posterior probability (GWPP)-based [18] confidence scoring.

In this method, the joint confidence of all component words in a recognition result is used to measure the confidence of a recognized utterance. The GWPP of a word is a measure of its correctness, or the probability of a binomial distributed "word correct" event. Thus, the probability of a "sentence correct" event is a product of all probabilities of component word correct events, assuming that all word events are statistically independent. A hypothesis with a probability of "sentence correct" event that is higher than a threshold is transferred as the final recognition result to the dialog processing module in the robot. More details about the GWPP word confidence module can be found in [18].

III. Experiments

A. Preparation of the modules

The acoustic models for adults were trained by using five hours dialogue speech from the ATR travel arrangement task database and 25 hours read speech of phonetically balanced sentences [19]. The training data was contaminated with eight types of noises listed in Table I at three types of SNR (20, 15, and 10dB), and the MMSE noise suppression was applied to the whole training data. A state-tying structure with 2,089 states was generated by using the MDL-SSS (Maximum Description Length Successive State Splitting) algorithm [20]. Each HMM state has five Gaussian distributions.

The feature vector consists of 12 MFCCs, Delta-pow, 12 Delta-MFCCs calculated with a 10-ms frame period and a 20-ms frame length. Cepstrum Mean Subtraction (CMS) was applied to the MFCC features, to reduce the effects of channel distortion.

The acoustic models for children were constructed by adaptation with an MLLR (Maximum Likelihood Linear Regression) algorithm using 12,000 words uttered by 238 child speakers in the CIAIR-VCV (Database of children's speech while playing video games) [21], which is provided from Nagoya university. The child training data was also contaminated with the same eight types of noises at 20, 15 and 10dB SNR.

For the MMSE noise suppression, we prepared a clean speech GMM with 512 Gaussian distributions using 24 log Mel-spectral energy coefficients. This clean speech GMM was trained with clean training data for adult speech only.

For the EPD module, 128-mixture GMMs were prepared using noisy speech and noise data. The feature vector is constituted by 12 MFCCs, Delta-pow and 12 Delta-MFCCs.

The language model is based on FSA (Finite State Automaton). The language model was constructed in order to recognize short Japanese sentences. This language model consists of 46 nodes and 205 links, and the lexicon size is 115.

The ATRASR speech recognition system developed by ATR Spoken Language Communication Laboratories was used in all experiments.

TABLE I
NOISE TYPES USED FOR TRAINING

Street	High-speed railway	Rice paddies
Airport lobby	Underground mall	Forest
Boiler room	Driving car	

B. Experimental conditions

To evaluate our ASR system's robustness to noise and to speakers of different ages, we tested it using a child and adult multichannel speech database that was recorded using the Robovie with the microphone array and a SANKEN CS-1, which is a directional microphone. This database consists of 1,464 short Japanese sentences uttered by 12 children and 12 adults. Each speaker uttered 61 sentences, such as "*Hello Robovie*", "*Come here*" and "*What can you do?*", in front of the Robovie. The children's ages ranged from 6 to 12. The distance between the speaker and the Robovie was 1 m.

We also recorded cafeteria noise using the same microphone array; this noise data was recorded at lunch time, therefore, it contains many types of noises such as talking voice, and clattery from dishes. The noise level was about 70 dBA. The cafeteria noise was recorded using the same volume level (amp gain) as when recording the speech database.

The speech data was contaminated by cafeteria noise at 65 dBA, 70 dBA, and 75 dBA.

C. Experimental Results

To investigate the performance of the individual techniques described in Section II, we evaluated several ASR systems with different structures, as listed in Table II.

1) Evaluation of robustness to noise:

Fig. 5 shows the word accuracies of several ASR systems for adults' speech that was contaminated by cafeteria noise at 65 to 75 dBA. The performance of system D, which uses the

TABLE II
EVALUATED STRUCTURES OF ASR SYSTEMS

System	A	B	C	D	E	F	G	H	I
Microphone type	1ch DPA	CS-1	array	array	array	array	array	array	array
Outlier-robust GSC	no	no	yes	yes	yes	yes	yes	yes	yes
MMSE noise suppression	no	no	no	yes	yes	yes	yes	yes	yes
Training data	clean	clean	clean	clean	noisy	noisy	noisy	noisy	noisy
AMs dependent on	adult	adult	adult	adult	adult	child	both	both	both
Segmentation	hand	hand	hand	hand	hand	hand	hand	auto	auto
GWPP-based rejection	no	no	no	no	no	no	no	no	yes

twelve-element microphone array, the RGSC, and the MMSE noise suppression, was better than that of system A, B and C. And, system E, which has acoustic models trained with noise-contaminated adults' speech, achieved the best performance. System E reduced the errors by 85.5 %, 84.3 %, 80.5 %, and 48.3 %, in comparison to system A, B, C and D respectively. Clearly, performance is widely improved by applying all individual techniques described in Section II.

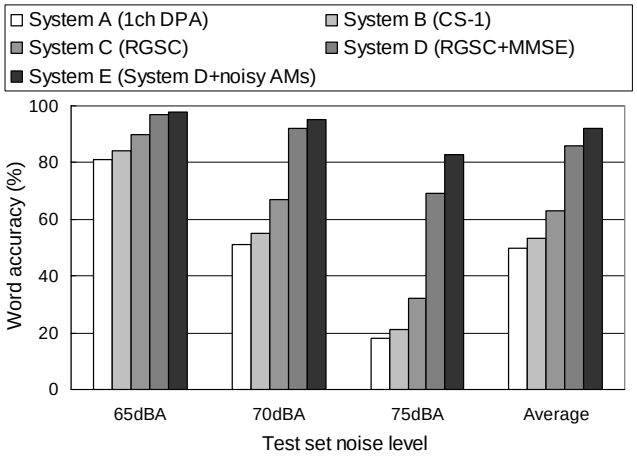

Fig. 5 Performance of system A to E for noise-contaminated adults' speech.

2) Evaluation of robustness to speakers of different ages:

Regarding the evaluation of robustness to speakers of different ages, we evaluated the use of hypothesis selection using AMs of both adults' and children's speech (System G). It contained two decoders with acoustic models depending on the age of the speaker (adult or child). For comparison, we evaluated the recognition performance of a system which contains acoustic models for adults' speech only (System E) and for children's speech only (System F).

Fig. 6 shows the word accuracies for adults' and children's speech. We can see that a model depending on the age of speaker which is matched to that of input speech achieved the best performance. The system using hypothesis selection (System G) performed almost equally to the systems in matched case. A slight degradation for adults' speech occurred when using adults AM only, as shown in the left part of Fig. 6. However, the right panel in Fig. 4 shows that the improvement for children speech is much more relevant.

3) Evaluation of the overall ASR system

We tested the recognition performance of System H, which includes the GMM-based EPD module. As evident from Fig. 6 the GMM-based EPD module introduced some errors because of misdetections of speech activity periods. Table III shows the word accuracies calculated only for the speech detected by the EPD module. Values in brackets are the word rejection rates by the EPD module. As can be observed from these results, the word accuracies with the EPD module (System H) is almost the same as when manual segmentation is used (System G).

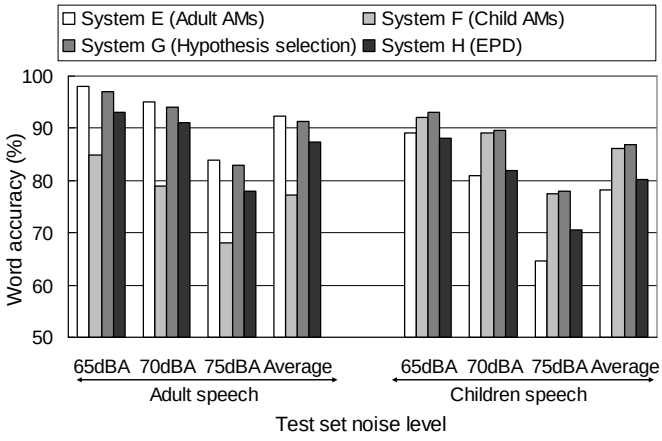

Fig. 6 Performance of system E to G for adults' and children's speech.

To improve the reliability of our ASR system, a GWPP-based confidence scoring module was implemented (System I). Table III shows the word accuracies obtained with the rejection module (EPD+GWPP). From these results, it is clear that word accuracies above 90 % were achieved at 70 dBA cafeteria noise. However a high rejection rate is also observed for high noise levels, indicating a tradeoff between word confidence and word rejection.

D. Evaluation of the overall ASR system in a real noisy environment

So far, the evaluation of the recognition system was realized by mixing clean speech and cafeteria noise, which were recorded separately. The purpose was to evaluate the robustness of the system at different signal-to-noise ratios. In this section, we provide a more realistic evaluation, by recording speech in a real noisy environment.

Eight adult speakers (four males and four females) uttered the same 61 short Japanese sentences of the previous experiments, resulting in a database of 488 utterances.

The robot was placed in the cafeteria, in the same conditions (location and lunch time) used to record noise data in the previous experiment. Also, the distance between the speaker and the Robovie was about 1 m.

TABLE III
WORD ACCURACIES (%) OF SYSTEM H AND I. VALUES IN PARENTHESIS ARE THE WORD REJECTION RATES (%) BY THE EPD AND THE GWPP MODULES.

	Adult speech			Children speech		
	65 dBA	70 dBA	75 dBA	65 dBA	70 dBA	75 dBA
System H (EPD)	95.84 (2.28)	94.87 (4.06)	89.52 (14.19)	90.42 (3.52)	86.96 (6.63)	80.78 (15.91)
System I (EPD+GWPP)	96.33 (3.14)	96.58 (6.49)	92.05 (19.12)	91.42 (5.70)	91.04 (13.46)	88.57 (28.63)

Recognition results indicated an average of 73 % word accuracy for all subjects. Word accuracies were between 70 to 84 % for seven subjects and 53 % for one of the subjects. Further analysis indicated that the SNR was about 10 dB for the seven subjects with higher scores and about 5 dB for the subject with the lowest score. The overall rejection rate (of correctly recognized words) was 8 %, while the overall rejection rate of insertions and incorrectly recognized words was 13 %.

A detailed analysis on the recognition errors revealed that some of the sentences, e.g. "utatteyo" ("sing!") and "tookudayo" ("it is far!"), showed low accuracies (less than 25 %). However it was observed that these errors were caused most due to rejection, rather than deletion or substitution. Also, monosyllabic sentences, like "hai" ("yes"), were found to be easier to be deleted, or misrecognized. In general, sentences composed by long lexicons were more reliably recognized.

Finally, although most of the evaluations in this paper was realized in off-line, the system was verified to run in real-time as well, by using a remote PC with a dual core Intel Xeon CPU at 3GHz each and 1GB RAM. The audio data from the twelve-element microphone array was sent from the Robovie to the remote PC by TCP/IP network transmission.

IV. Conclusion and Future Works

In this paper, we described a robust ASR system for communication robots, and evaluated its robustness to real noisy environments and to speakers of different ages. In our ASR system, a twelve-element microphone array arranged in the robot chest, an RGSC-based microphone array processing, an MMSE-based feature-space noise suppression, and multi-conditionally trained acoustic models were used to improve robustness. Moreover, to improve the robustness to speakers of different ages, we used two decoders for children's and adults' speech respectively. Finally, the recognition results were scored using GWPP-based confidence measure, for reducing insertion errors. Experimental results in several noise level conditions indicated that our ASR system could achieve word accuracies of more than 80 % with 70 dBA of background cafeteria noise, for both adult and children speech. Further evaluation in a real noisy environment resulted in 73 % word accuracy for adult speech.

As next steps of our work, a dialogue module will be developed for evaluating human-robot interaction in a real environment. We also intend to include a paralinguistic information extraction module [14], for also allowing a non-verbal communication between the robot and a human.

Further, sound source localization techniques can be used for adaptation of the main lobe of the beamformer, since currently the robot assumes that the speaker is in front of its chest.

Acknowledgment

This work was partly supported by the Ministry of Internal Affairs and Communications.

References

[1] C. Breazeal and B. Scassellati, A context-dependent attention system for a social robot, *Int. Joint Conf. on Artificial Intelligence(IJCAI'99)*, pp. 1146-1151, 1999.

[2] K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, and H. Kitano: 'Real-Time Auditory and Visual Multiple-Object Tracking for Robots,' Proc. Int. Joint Conf. on Artificial Intelligence, pp.1425-1432, 2001.

[3] O. Sugiyama, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, "Three-layered Draw-Attention Model for Humanoid Robots with Gestures and Verbal Cues," IEEE/RSJ Int. Conf. on Intelligent robots and systems (IROS2005), pp. 2140-2145, 2005.

[4] T. Shibata, "An overview of human interactive robots for psychological enrichment", *Proceedings of the IEEE*, Vol.92, No.11, 2004.

[5] W. Burgard, A. B. Cremers, D. Fox, D. Hahnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun, The interactive museum tour-guide robot, *National Conference on Artificial Intelligence*, pp. 11-18, 1998.

[6] M. Fujita, AIBO; towards the era of digital creatures, *Int. J. of Robotics Research*, Vol. 20, No. 10, pp. 781-794, 2001.

[7] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial, *Human Computer Interaction*, Vol. 19, No. 1-2, pp. 61-84, 2004.

[8] M. Shiomi, T. Kanda, H. Ishiguro, and N. Hagita, Interactive Humanoid Robots for a Science Museum, 1st Annual Conference on Human-Robot Interaction (HRI2006), 2006.

[9] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Kitano, Applying Scattering Theory to Robot Audition System, IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2003), pp. 1147-1152, 2003.

[10] Asoh, H., Hayamizu, S., Hara, I., Motomura, Y., Akaho, S., and Matsui, T. "Socially Embedded Learning of the Office-Conversant Mobile Robot Jijo-2," Int. Joint Conf. on Artificial Intelligence (IJCAI), 1997.

[11] T. Takatani, S. Ukai, T. Nishikawa, H. Saruwatari, and K. Shikano, "Blind sound scene decomposition for robot audition using SIMO-model-based ICA," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2005), pp. 215-220, 2005.

[12] Y. Ohashi, T. Nishikawa, H. Saruwatari, A. Lee, and K. Shikano, "Noise-robust hands-free speech recognition based on spatial subtraction array and known noise superimposition," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2005), pp. 533-537, 2005.

[13] S. Yamamoto, K. Nakadai, J.M. Valin, J. Rouat, F. Michaud, K. Komatani, T. Ogata, and H. Okuno, "Making a robot recognize three simultaneous sentences in real-time," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2005), pp. 897-902, 2005.

[14] C. T. Ishi, H. Ishiguro, N. Hagita: "Evaluation of prosodic and voice quality features on automatic extraction of paralinguistic information," accepted to *IROS 2006*.

[15] W. Herbordt, T. Horiuchi, M. Fujimoto, T. Jitsuhiro, and S. Nakamura, "Hands-free speech recognition and communication on PDAs using microphone array technology," Proc. *ASRU*, pp. 302-307, 2005.

[16] W. Herbordt, H.Buchner, S.Nakamura, and W.Kellermann, "Application of a double-talk resilient DFT-domain adaptive filter for bin-wise stepsize controls to adaptive beamforming," Proc.*IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, pp. 175.181, May 2005.

[17] S. Matsuda, T. Jitsuhiro, K. Markov, and S. Nakamura, "ATR Parallel Decoding Based Speech Recognition System Robust to Noise and Speaking Styles," *IEICE Trans. Inf. & Syst.*, vol. E89-D, No. 3, pp. 989--997, 2006.

[18] F.K. Soong, W.K. Lo, and S. Nakamura, "Generalized Word Posterior Probability (GWPP) for Measuring Reliability of Recognized Words," Proc. *SWIM2004*, 2004

[19] T. Takezawa, T. Morimoto, and Y. Sagisaka, "Speech and language databases for speech translation research in ATR," In Proc. *the 1st International Workshop on East-Asian Language Resources and Evaluation (EALREW 98)*, pp. 148--155, 1998.

[20] T. Jitsuhiro, T. Matsui, and S. Nakamura, "Automatic Generation of Non-uniform HMM Topologies Based on the MDL Criterion," *IEICE Trans. Inf. & Syst.*, vol. E87-D, no. 8, pp. 2121--2129, 2004.

[21] Center for Integrated Acoustic Information Research, http://db.ciair.coe.nagoya-u.ac.jp/eng/dbciair/dbciair2/kodomo.htm