Proceedings of 2004 IEEE/RSJ International Conference on
Intelligent Robots and Systems
September 28 - October 2, 2004, Sendai, Japan

Assessment of General Applicability of Robot Audition System by Recognizing Three Simultaneous Speeches

Shun'ichi Yamamoto†, Kazuhiro Nakadai‡, Hiroshi Tsujino‡, and Hiroshi G. Okuno†

†Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
shunichi@kuis.kyoto-u.ac.jp, okuno@i.kyoto-u.ac.jp

‡Honda Research Institute Japan, Co., Ltd.
8-1 Honcho, Wako-shi, Saitama 351-0114, Japan
{nakadai, tsujino}@jp.honda-ri.com

Abstract— **Robot audition is a critical technology in creating an intelligent robot operating in daily environments. We have developed such a robot audition system by using a new interface between sound source separation and automatic speech recognition (ASR). A mixture of speeches captured with a pair of microphones installed in the ear positions of a humanoid is separated into each speech by using active direction-pass filter (ADPF). The ADPF extracts a sound source originating from a specific direction in real-time by using interaural phase and intensity differences. The separated speech is recognized by a speech recognizer based on the missing feature theory (MFT). By using a missing feature mask, the MFT based ASR neglects distorted and missing features caused during the speech separation. A missing feature mask for each separated speech is generated in speech separation and is sent to the ASR with the separated speech. Thus, this new integration improves the performance of ASR. However, the generality of this robot audition system has not been assessed so far. In this paper, we assess its general applicability by implementing it on the three humanoids, i.e., ASIMO of Honda, SIG2, and Replie of Kyoto University. By using three simultaneous speeches as benchmarks, the robot audition system improved the performance of ASR over 50% in every humanoid, and thus its general applicability was confirmed.**

I. Introduction

Robot audition is one of the most important technologies in order to interact with people as human partners in the near feature. A robot should have the capability of social interaction with people. One of the most important functions to achieve this is verbal communication. In addition, the robot should have the capability to pay attention to specific sound events such as environmental sounds, spoken language, and music.

To achieve such robot audition, the robot should handle a mixture of sounds, because humans and robots usually hear a mixture of sounds, not sounds of a single source. Some human-robot communication systems use a microphone attached close to each talker's mouth in order to avoid handling a mixture of sounds [1], [2]. However, in real environments, a robot should handle a mixture of sounds captured by its own microphones. The three basic functions to handle a mixture of sounds are sound source localization, separation and recognition. In robot audition, Nakadai *et al.* used psychological clues for binaural hearing [3]. Several studies on a microphone array have been reported; for example, Asano *et al.* used beamforming techniques for the "Jijo-2" robot [4].

Since robot audition ranges from signal to conceptual level, a hierarchical structure of auditory processing is required. At the signal level, three basic functions, that is, sound source localization, separation, and recognition, are needed. To climb up to the conceptual level, signal-to-symbol transformation is required. Symbols may be represented by text or ontologies. Speech is recognized as text by automatic speech recognition (ASR). Environmental sounds may be recognized as either names of auditory events or sound-imitation words, i.e., onomatopoeia [5].

Since human communication is performed mainly by spoken languages, we focus on the interface between sound source separation and ASR. Usually, the ASR community focuses on robust ASR, assuming that one person speaks under noisy environments. For example, ASIMO of Honda [6], QRIO of Sony [7], Kismet of the MIT AI Lab [8], and ROBITA of Waseda University [2] can interact with people by recognizing speech and gestures. All four robots can localize a sound source, but assume there is one talker for ASR. The latter two robots use a separate microphone attached near the mouth of each speaker.

In this paper, we focus on a noise-robust automatic speech recognition to achieve robot audition for the robot. We adopt the *Missing Feature Theory* (MFT) to design the interface between sound source separation and ASR to improve the robustness against dynamic noises and simultaneous speeches [9]. Since robot audition is usually dedicated to a particular robot, its general applicability to other robots is not assessed. We have been assessed the generality of our robot audition system to apply it to two different humanoid robots, i.e., SIG2, and Replie of Kyoto University. However, these robots are similar in many characteristics. They have soft skin, the human-shaped silicon ears located at human's ear positions, and a pair of microphones installed at external auditory meatuses in the ears. In addition, the shapes of these robots' head are similar to that of human. Therefore, we will assess the generality of our auditory system by applying it to a robot whose characteristics are quite different from those of SIG2 and Replie. We use ASIMO of Honda which has hard skin, a angular face, and microphones at different positions.

0-7803-8463-6/04/$20.00 ©2004 IEEE

II. Speech Recognition for Robot Audition

To have the capability of robot audition, a robot would need to cope with the following difficult situations.

- A robot needs to be able to listen to a specific sound source in noisy environments. This capability in humans is known as the "cocktail party effect".
- A robot should be able to listen to several speeches simultaneously. This is required to cope with the case that someone or something playing sounds interrupts a conversation. It is known as "barge-in" in spoken dialog systems.

The improvement of robustness against noises in ASR has been studied extensively, in particular, in the AURORA project [10]. One method for noise-robust ASR is *multi-condition training*, that is, training on a mixture of clean speech and noises [11], [12]. This is currently the most common method for car and telephone applications. Since an acoustic model obtained by multi-condition training reflects all expected noises in specific conditions, ASR using the acoustic model is effective as long as trained and static noises are dominant. This is also effective in robots.

Nakadai *et al.* developed the interface between a sound source separation system and ASR, and demonstrated that their system can recognize three simultaneous speakers with high accuracy [13]. The sound source separation system called *active direction-pass filter* (ADPF)[14] separated sound sources by using directional information given by visual and or auditory processing. Since the spectral features of a separated sound are severely distorted, they use direction- and speaker-dependent acoustic models for ASR at the same time, and choose the most appropriate recognition result. This is a brute-force approach and expensive in computational resources. For the moment, no other system that can recognize three simultaneous speakers has been reported in the literature.

The issues with their interface between sound source separation and ASR in realizing robot audition are summarized as follows.

1) Assessment of general applicability to other humanoids.
2) Requirement of direction- and speaker-dependent acoustic models.
3) Expensive computing resources and slow processing time.

For the first issue, since they use only the upper-torso humanoid *SIG*, the generality of their methodology has not been evaluated so far. In this paper, we will assess the generality of our MFT-based ASR by using three different humanoid test-beds.

The second issue indicates that the system has difficulties in coping with an unknown speaker or sounds originating from an unexpected direction. The performance is low under dynamically changing noisy environments and different acoustic environments, since each direction- and speaker-dependent acoustic model is tuned for a particular environment by multi-condition training.

The last issue is related to the second one. When three speakers may place one of 10-degree wise position, their system needs to exploit 51 combinations of direction- and speaker-dependent acoustic models. In other words, 51 ASRs with each combination of acoustic model runs in parallel against each separated speech. This requires a lot of computational resources, and thus is not suitable for autonomous robots whose physical body size is limited.

Our idea is to adopt the MFT to design the interface between ADPF and ASR to cope with the above issues and achieve robot audition under daily environments where noises change dynamically. The proposed ASR will run with a *single*-direction- and speaker-independent acoustic model with clean-condition training. Therefore, its computational resource is the same as normal ASR's.

III. MFT-Based Interface

An MFT-based ASR has been studied as a promising way to improve the robustness of ASR [15], [16], although most studies have been done in off-line and simulated environments. In this method, spectral subbands distorted by noises are detected from input speech as missing features. The detected missing features are masked on recognition so as not to affect the system badly. Therefore, this method is more flexible when noises change dynamically and drastically. In this paper, we use the MFT to interface

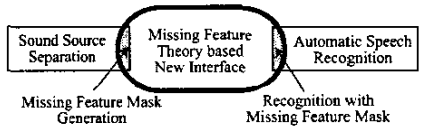


Fig. 1. New Interface based on Missing Feature Theory

sound source separation with ASR. A model of the MFT-based interface is illustrated in Figure 1. This interface uses a *missing feature mask* to avoid deterioration of speech recognition caused by the missing features. To introduce the missing feature mask, modifications in sound source separation and ASR are necessary as shown in the dark areas in Figure 1. In sound source separation, the missing feature mask is generated from missing features that are detected by comparing a separated speech with a target speech. In ASR, a speech recognizer is modified to make use of the missing feature mask. The following sections describe these modifications in detail.

A. ASR with Missing Features

MFT-based ASR is based on the Hidden Markov Model (HMM) by using Mel-Frequency Cepstrum Coefficients (MFCC) and the Hidden Markov Model (HMM), which is a common feature vector in normal ASR systems. Since the missing feature mask is introduced in MFT-based ASR, the speech recognizer is modified to have one more input of missing feature mask input (as is shown in Figure 2 in the section Robot Audition System with MFT-Based Interface). The missing feature mask is generated in sound source separation as described in the next section. The

input feature vectors are masked by the missing feature mask in a speech recognizer. In normal ASR systems, estimation of a path with maximum likelihood is based on state transition probabilities and output probability in the Viterbi algorithm. MFT-based ASR uses a different estimation of the output probability, which is specified as follows.

Let $o(x|S)$ be the output probability of feature vector x in state S. The output probability is defined by

$$o(x|S) = \sum_{k=1}^{K} P(k|S) \exp\left\{ \sum_{l=1}^{L} m_l \log f(x_l|k,S) \right\},$$

where K is the number of Gaussian mixture, $f(x_l|k,S)$ is the probability density function of Gaussian distribution, L is the size of vector $x = (x_1, x_2, \cdots, x_L)$, and vector $m = (m_1, m_2, \cdots, m_L)$ is mask vector. The mask vector is defined in the next section.

This equation means that only reliable features are used in the probability calculation. Therefore, the recognizer can avoid severe degradation of performance caused by unreliable features.

B. Missing Feature Mask Generation in Sound Source Separation

For sound source separation, we use an ADPF, which extracts sound originating from the specified direction, by using a pair of microphones. The detailed algorithm of ADPF is described in [13]. ADPF first extracts an *interaural phase difference (IPD)* and an *interaural intensity difference (IID)* for each subband. Then, it estimates the sound source direction from IPD and IID by scattering theory [17]. Since scattering theory provides an accurate estimation of IPD and IID for a spherical robot head with two microphones, we need not measure the head-related transfer function for each acoustic environment [14].

After estimation of IPD and IID, ADPF selects the pass range according to the pass range function. The pass range function specifies a narrow pass range for the front direction due to ADPF's high sensitivity, while it specifies a wider one for the peripherals due to ADPF's low sensitivity. Practically, the pass range is $\pm 10°$ for the sound source direction of $30°$, and $\pm 5°$ for that of $0°$. Finally, ADPF collects subbands of input whose IPD and IID are within the pass range. This collection is treated as a separated speech.

MFCC of separated speech are calculated directly from the collected subbands. A missing feature mask is generated by comparing MFCC of the separated speech with that of the corresponding clean speech. This kind of mask is called an *a priori mask*, because mask generation heuristics use information about corresponding clean speech [18].

Finally, the missing feature mask is obtained as a matrix of an MFCC vector and time frame. Each value in the matrix is a belief factor that represents whether the corresponding value in the MFCC vectors obtained by the input signal is reliable or not. The belief factor can be a continuous value from 0 to 1, or can be a discrete value of 0 or 1. The latter is used in this paper.

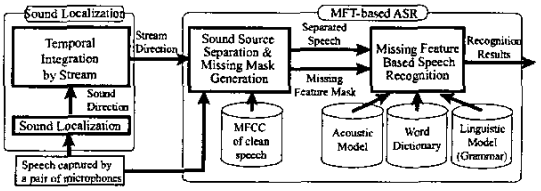


Fig. 2. Robot Audition System with MFT-based Interface

The detailed algorithm is specified as follows:

1) Let X and Y be feature vectors of captured speech and the corresponding clean speech, respectively. The feature vector consists of 26 features with 12 MFCCs, power of signal, 12 Δ MFCCs, and Δ power. In recognition, the feature of power is not used.
2) Let $M_k(i)$ be the mask value of the ith feature in the kth frame. $M_k(i)$ is obtained by

$$M_k(i) = \begin{cases} 1 & \text{if } |X_k(i) - Y_k(i)| < T, \\ 0 & \text{otherwise.} \end{cases}$$

where T is an experimentally obtained threshold.
3) $\Delta M_k(i)$ is defined by

$$\Delta M_k(i) = M_{k-2}(i) M_{k-1}(i) M_{k+1}(i) M_{k+2}(i).$$

4) Thus, the mask vector m of the kth frame is $(M_k(1), \cdots, M_k(13), \Delta M_k(1), \cdots, \Delta M_k(13))$.

IV. Robot Audition System with MFT-Based Interface

We implemented a robot audition system by using the new interface between sound source separation and ASR based on the MFT. The architecture of the system is shown in Figure 2. It consists of sound source localization and MFT-based speech recognition subsystems.

A. Sound Source Localization Subsystem

The sound source localization subsystem localizes multiple sound sources captured by microphones embedded in a robot. The detailed algorithm of sound source localization is described in [14]. The sound source localization module extracts local peaks from the left and right power spectrums and clusters a harmonic sound according to harmonic relationships. Then it calculates IPD and IID of the peaks included in the extracted harmonic sound and calculates distances between the results and IPD and IID hypotheses created by the scattering theory for each sound direction. The calculated distances are transferred to belief factors on IPD and IID. The belief factors on IID and IPD are integrated based on the Dempster-Shafer theory [19] to get robust sound localization in the real world. As a result of the integration, a direction with maximum value is regarded as that of the sound source. The sound source is localized in a horizontal plane by using a pair of microphones that are installed in the left and right ear positions of a humanoid.

The temporal integration module forms a sound stream as a temporal sequence of sound localization events by using a Kalman filter. The sound stream provides accurate

and robust sound direction information for sound source separation because the Kalman filter reduces measurement and process noises in localization.

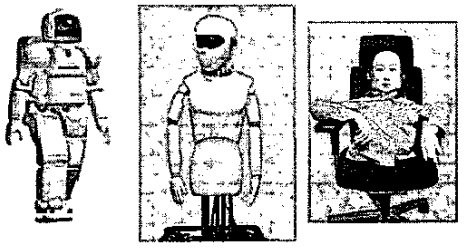
B. MFT-based Speech Recognition Subsystem

The MFT-based speech recognition subsystem is based on the new interface described in the previous section, and recognizes noisy speech such as simultaneous speeches. We use the CASA Toolkit (CTK) [15] based on MFT as a speech recognizer. The CTK can use monophones and triphones for an acoustic model, while CTK currently does not support use of statistical language models. The ADPF separates multiple sound sources by using a stream direction obtained by a sound source localization subsystem and a captured sound mixture, and estimates a missing feature mask by comparing a speech separated by the ADPF with a target clean speech. The CTK performs isolated word recognition against the separated speech by using an acoustic model, a word dictionary, grammar, and the missing feature mask.

V. Evaluation

We performed three experiments to confirm the general capabilities and improvements by introducing the MFT. The three different humanoids ASIMO, *SIG2*, and *Replie* are used for experiments. The humanoids ASIMO, *SIG2*, and *Replie* are shown in Figures 3a) – c), respectively. In these experiments, our system works on Pentium 4 2.53 GHz PC running Linux, and binaural sounds captured by each humanoid are processed off-line.

A. Humanoids for Test-beds



a) ASIMO (Honda) b) *SIG2* (Kyoto Univ.) c) *Replie* (Kyoto Univ.)

Fig. 3. Humanoid Robots

Fig. 4. Ears in Humanoid Test-beds

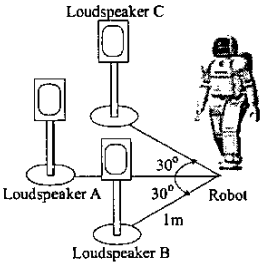


Fig. 5. Experiment

SIG2 and *Replie* have soft skin and human-shaped ears made of silicon shown in Figure 4. Their microphones are installed in external auditory meatuses in the ears located at human ear positions. The pinnae in the ears improves the front directivity of 10 dB. The appearances of these two humanoids have some differences, because *SIG2* was designed by a professional designer in consideration of aesthetic appearance, while *Replie* was made by molding a Japanese woman and has a full body although only the upper half of the body is shown in Figure 3b). Thus, the acoustics of these humanoids have some differences although many features between the two humanoids are common.

On the other hand, ASIMO has a hard cover, an angular face and a pair of microphones at the different positions from *SIG2* and *Replie*'s ones. Therefore the acoustics are quite different from the other two humanoids.

B. Acoustic Model for Speech Recognition

Only one HMM-based acoustic model trained on clean speech is used for recognition of separated speech. The training data includes a total of 25 male and female speakers' utterance sets. Each utterance set consists of 216 phonemically-balanced Japanese words. The feature vector of the acoustic model has a dimension of 25 (12 MFCC + 12 Δ MFCC + Δ power). The number of states and mixtures in HMM are 3 and 8, respectively.

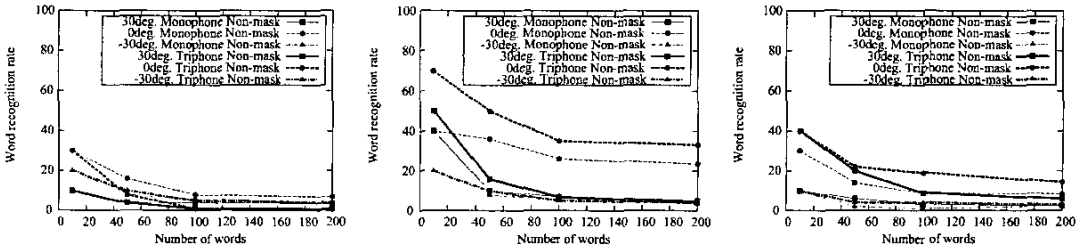
C. Experiments

The robot audition system is evaluated by recognizing three simultaneous speeches, as shown in Figure 5. Three loudspeakers located at fixed directions of 0° and ±30° are used for sound sources. The distance between the loudspeakers and the robots is 1 m. The rooms are $7.5\,m \times 9\,m$ with 0.5 sec of reverberation time for ASIMO, and $5\,m \times 4\,m$ with 0.35 sec of reverberation time for *SIG2* and *Replie*. The loudspeakers play 200 combinations of three different words selected from a set of 216 phonemically-balanced Japanese words. The humanoids capture the mixture of acoustic signals of these three words and omnidirectional background noises. The acoustic signal of each word is extracted by the ADPF and recognized by isolated word recognition. ADPF is given the radias of spherical robot head, and the left and right ear positions as parameters for scattering theory.

The isolated word recognition is evaluated by using two metrics: a word recognition rate whereby target words are categorized into the target category correctly, and a mis-categorized rate whereby non-target words are mis-categorized into the target category. Several conditions are changed in the experiments, which are as follows:

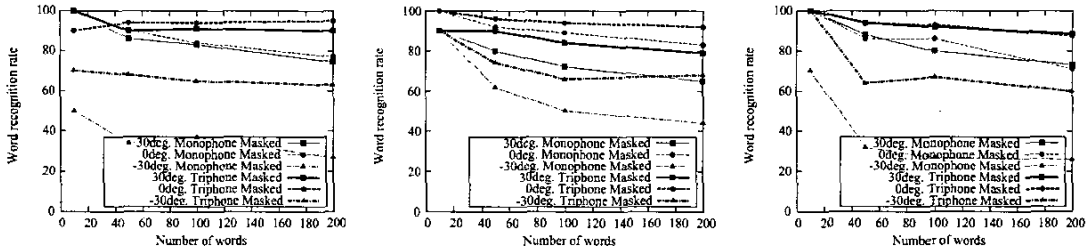
- word dictionary size – 10, 50, 100, and 200 words,
- type of acoustic model – monophone and triphone,
- missing feature mask – used and unused, and
- robots – ASIMO, *SIG2*, and *Replie*.

The word recognition rates by using these humanoids are shown in Figures 6 and 7. The isolated word recognition rates without using the missing feature mask in ASIMO,



a) ASIMO b) *SIG2* c) *Replie*

Fig. 6. Isolated Word Recognition Rates *without* Missing Feature Mask



a) ASIMO b) *SIG2* c) *Replie*

Fig. 7. Isolated Word Recognition Rates *with* Missing Feature Mask

TABLE I
MIS-CATEGORIZED RATES *with* MISSING FEATURE MASK (%)

	monophone				triphone			
# of words	10	50	100	200	10	50	100	200
ASIMO	19.4	5.8	2.4	1.3	10.2	4.1	1.9	0.7
SIG2	5.2	2.2	0.8	0.7	3.0	1.8	0.8	0.3
Replie	15.5	4.8	2.2	1.2	4.4	2.7	0.8	0.2

SIG2 and *Replie*, are shown in Figures 6a), b), and c) respectively. Those using the missing feature mask in these humanoids are shown in Figures 7a) – c). When the missing feature mask is not used, a separated speech is recognized by assuming that every feature is reliable. This means that the result is the same as the recognition of separated speech by normal ASR systems. Each figure includes results of left (30°), center (0°) and right (−30°) speakers by using the monophone (thin lines) and triphone (thick lines) as acoustic models. The x and y axes indicate the number of words included in the word dictionary and the isolated word recognition rate, respectively.

Generally, as the number of words in the word dictionary increases , word recognition rates decrease. This tendency is remarkable in the results obtained without using the missing feature mask. On the other hand, the results obtained using the missing feature mask show robustness against the increase in the number of words. The word recognition rates improve over 80% in the case of a 200-word dictionary as well. In this case, the mis-categorized rates are less than 5 % in every robot when a triphone are used as an acoustic model. The mis-categorized rates with the missing feature mask is shown in Table I. This demonstrates that the MFT-based approach is efficient for speech recognition in robots.

When monophone and triphone are compared in each figure, the results using triphone are about 10% better than those using monophone. Actually, triphone-based speech recognition is common in normal ASR systems. This shows that triphone is effective in missing feature based ASR for robots. When both the triphone and the missing feature mask are used, the recognition rate reaches about 90% even in the case of a 200 word dictionary. The performance exceeds that of another approach that integrates multiple speech recognition results obtained by using multiple direction and speaker-dependent acoustic models [13]. This demonstrates the efficiency of the proposed robot audition system in terms of performance as well as processing speed, because only a single acoustic model is used for speech recognition.

The MFT-based ASR works well in all three humanoids – ASIMO, *SIG2* and *Replie* as shown in Figure 7. These humanoids have different heads and bodies. Considering that auditory processing is sensitive to a small change in the acoustic environment, their performances are expected to be different, but they have similar performance. This indicates the generality of the robot audition system, especially, the new interface between sound source separation and ASR.

VI. DISCUSSION AND FUTURE WORK

The auditory awareness system based on the proposed interface between sound source separation and ASR improves the performance of the system and processing speed in the recognition of three simultaneous speeches. The works that reported on simultaneous speech recognition [20], [21], [13] treated sound source separation and ASR independently. In other words, sound source separation

is treated as a simple preprocessor of ASR, and thus, the performance was not so good. Because the MFT-based interface takes such characteristics into account, the performance is better. Therefore, we can say that the MFT-based interface is more suitable for robot audition. To improve the performance of the system, a combination of multi-condition training and the MFT could be effective.

We demonstrated the generality of the robot audition system through the application of the three humanoids in the rooms with different acoustic conditions. Usually, in robotic research, methodologies applied to a robot are dedicated to the robot and a specific environment. Because of this, the generality of the methodologies have not been evaluated so much, although it is frequently considered. Thus, the evaluation of generality is one of the most important issues from the viewpoint of applicability of the method.

In this paper, the system uses the a priori missing feature mask that is obtained from clean speech as a template. This is good for recognition of a specific word. Taking a more general situation into consideration, detection of missing feature without using clean speech as a template is necessary. This is a challenging future work.

In this experiment our system recognized three simultaneous speech well. The number of simultaneously recognizable speakers depends on the performance of sound source separation. The quality of separated speech is good when the number of simultaneous speakers is less than or equal three. In another experiment the performance of our system is lower than in this experiment when speakers are located at intervals of less than 30°. Therefore 30° is the limitation of our method.

The system performs well in isolated word recognition. This is good for simple dialog systems, but it is not enough to recognize longer sentences in conversation, complex dialog and dictation. On the other hand, we consider that complete recognition of sentences such as dictation is difficult. Keyword recognition by using word spotting techniques will be the first step to solve this problem.

VII. Conclusion

Robot audition is a critical technology in creating an intelligent robot that operates in daily environments. To achieve such robot audition, we reported the MFT-based interface between sound source separation and ASR in this paper. The robot audition system based on the interface exhibited high performance in the recognition of three simultaneous speeches. This means that the MFT is a promising method to improve the performance in systems that treat sound source separation and ASR separately. In addition, the generality of the interface is assessed through application to the more different humanoid robot, ASIMO as well as *SIG2*, and *Replie*.

Acknowledgment

We thank to Prof. Phil Green and Dr. John Barker, Sheffield University for their help on CTK and valuable discussions. We also thank to Prof. Tatsuya Kawahara and Dr. Hiroaki Nanjyo, Kyoto Univ. and Dr. Takahiro Miyashita, ATR for their discussions.

References

[1] C. Breazeal, "Emotive qualities in robot speech," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2001)*. IEEE, 2001, pp. 1389–1394.

[2] Y. Matsusaka, T. Tojo, S. Kuota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano, and T. Kobayashi, "Multi-person conversation via multi-modal interface — a robot who communicates with multi-user," in *Proc. of Eurospeech-1999*. 1999, pp. 1723–1726.

[3] K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, and H. Kitano, "Real-time auditory and visual multiple-object tracking for robots," in *Proc. of IJCAI-2001*, 2001, pp. 1424–1432.

[4] F. Asano, H. Asoh, and T. Matsui, "Sound source localization and signal separation for office robot "Jijo-2"," in *Proc. of IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI-99)*, 1999, pp. 243–248.

[5] K. Ishihara, Y. Tsubota, and H. Okuno, "Automatic transformation of environmental sounds into sound-imitation words based on Japanese syllable structure," in *Proc. of Eurospeech-2003*. ESCA, 2003, pp. 3185–3188.

[6] Y. Sakagami, R. Watanabe, C. Aoyama, S. Matsunaga, N. Higaki, and K. Fujimura, "The intelligent ASIMO: System overview and integration," in *Proc. of IROS-02*. IEEE, 2002, pp. 2478–2483.

[7] Y. Kuroki, M. Fujita, T. Ishida, K. Nagasaka, and J. Yamaguchi, "A small biped entertainment robot exploring attractive applications," in *Proc. of ICRA-2003*. IEEE, 2003, pp. 471–476.

[8] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," in *Proc. of IJCAI-99*, 1999, pp. 1146–1151.

[9] S. Yamamoto, K. Nakadai, H. Tsujino, T. Yokoyama, and H. G. Okuno, "Improvement of robot audition by interfacing sound source separation and automatic speech recognition with missing feature theory," in *Proc. of ICRA-2004)*. IEEE, 2004, pp.1517–1523.

[10] D. Pearce, "Developing the ETSI AURORA advanced distributed speech recognition front-end & what next," in *Proc. of Eurospeech-2001*. ESCA, 2001.

[11] R. P. Lippmann, E. A. Martin, and D. B. Paul, "Multi-styletraining for robust isolated-word speech recognition," in *Proc. of ICASSP-87*. IEEE, 1987, pp. 705–708.

[12] M. Blanchet, J. Boudy, and P. Lockwood, "Environmentadaptation for speech recognition in noise," in *Proc. of EUSIPCO-92*, vol. VI, 1992, pp. 391–394.

[13] K. Nakadai, H. G. Okuno, and H. Kitano, "Robot recognizes three simultaneous speech by active audition," in *Proc. of ICRA-2003*. IEEE, 2003, pp. 398–403.

[14] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Kitano, "Applying scattering theory to robot audition system: Robust sound source localization and extraction," in *Proc. of IROS-2003*. IEEE, 2003, pp. 1157–1162.

[15] J. Barker, M. Cooke, and P. Green, "Robust asr based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise," in *Proc. of Eurospeech-2001*. ESCA, 2001, pp. 213–216.

[16] P. Renevey, R. Vetter, and J. Kraus, "Robust speech recognition using missing feature theory and vector quantization," in *Proc. of Eurospeech-2001*. ESCA, 2001, pp. 1107–1110.

[17] J. J. Bowman, T. B. A. Senior, and P. L. E. Uslenghi, *Electromagnetic and Acoustic Scattering by Simple Shapes*. Hemisphere Publishing Co., 1987.

[18] K. Palomaki, G. Brown, and J. Barker, "Missing data speech recognition in reverberant conditions," in *Proc. of ICASSP-2002*. IEEE, 2002, pp. 65–68.

[19] A. Dempster, "Upper and lower probabilities induced by a multi-valued mapping," *Annals of Mathematical Statistics*, vol. 38, pp. 325–339, 1967.

[20] H. G. Okuno, T. Nakatani, and T. Kawabata, "Understanding three simultaneous speakers," in *Proc. of IJCAI-1997*, 1997, pp. 30–35.

[21] A. Koutras, E. Dermatas, and G. Kokkinakis, "Improving simultaneous speech recognition in real-room environments using overdetermined blind source separation," in *Proc. of Eurospeech-2001*, 2001, pp. 1009–1012.