DESIGN AND IMPLEMENTATION OF A ROBOT AUDITION SYSTEM FOR AUTOMATIC SPEECH RECOGNITION OF SIMULTANEOUS SPEECH

Shun'ichi Yamamoto[*], *Kazuhiro Nakadai*[†], *Mikio Nakano*[†], *Hiroshi Tsujino*[†],
Jean-Marc Valin[‡], *Kazunori Komatani*[*], *Tetsuya Ogata*[*], *and Hiroshi G. Okuno*[*]

* Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan

†Honda Research Institute Japan Co., Ltd., 8 - 1 Honcho, Wako-shi, Saitama 351-0114, Japan

‡CSIRO ICT Centre, Cnr Vimiera & Pembroke Rds, Marsfield NSW 2122, Australia

ABSTRACT

This paper addresses robot audition that can cope with speech that has a low signal-to-noise ratio (SNR) in real time by using robot-embedded microphones. To cope with such a noise, we exploited two key ideas; *Preprocessing* consisting of sound source localization and separation with a microphone array, and system integration based on *missing feature theory (MFT)*. Preprocessing improves the SNR of a target sound signal using geometric source separation with multi-channel post-filter. MFT uses only reliable acoustic features in speech recognition and masks unreliable parts caused by errors in preprocessing. MFT thus provides smooth integration between preprocessing and automatic speech recognition. A real-time robot audition system based on these two key ideas is constructed for Honda ASIMO and Humanoid SIG2 with 8-ch microphone arrays. The paper also reports the improvement of ASR performance by using two and three simultaneous speech signals.

Index Terms— Robot audition, missing feature theory, geometric source separation, automatic speech recognition

1. INTRODUCTION

Robots should listen to their surrounding world by *their own ears (microphones)* to recognize and understand the auditory environments. We call this kind of artificial listening capability "robot audition". It has been studied to improve real-time auditory processing in the real world for the past five years at robotics-related conferences. Robot audition is considered as an essential function to understand the surrounding auditory world such as human voices, music, and other environmental sounds. One good example of behavioral intelligence in robot audition is active audition [1] which improves robot audition by integrating it with active motion such as turning to and approaching a target sound source, and asking the user again what the robot failed to listen to. This means behavioral intelligence is essential for robot audition, because selection of an appropriate behavior for better robot audition depends on where the robot is located, and therefore, it requires high intelligence. The ultimate goal in robot audition is real-time automatic speech recognition (ASR) under noisy and reverberant environments. To cope with such a noisy speech signal, noise adaptation techniques such as *multi-condition* training [2] and *Maximum-Likelihood Linear Regression (MLLR)* [3] are commonly used. Because these techniques can deal with the trained noises well, they are used for telephony applications and for car navigation systems. However, a robot should recognize several things simultaneously because multiple sound sources exist simultaneously. In addition, input signals to microphones embedded in robots inevitably include various kinds of noise such as robot motor noise, environmental noise, and room reverberation. Since the signal-to-noise ratio (SNR) of input signals is extremely low and noises are not always known in advance, common techniques are in general unsuitable for robot audition. To solve this problem, we exploited the following two key ideas:

1. *Preprocessing of ASR* such as sound source localization and separation using a robot-embedded microphone array.
2. *Missing Feature Theory (MFT)* [4, 5] that integrates preprocessing with ASR by masking unreliable features included in preprocessed signals and using only reliable features for recognition.

We implemented a real-time robot audition system for Honda ASIMO and Humanoid SIG2 with 8-ch microphone arrays. The system was evaluated in terms of recognition of single and simultaneous speech when a robot noise was present.

The rest of this paper is organized as follows: Section II explains our key ideas for robot audition with related work. Section III describes the implementation of our robot audition system based on the approaches. Section IV evaluates our system. The last section concludes this paper.

2. KEY IDEAS

This section describes our two key ideas for achieving robot audition – preprocessing and missing-feature-theory-based integration. When the system recognizes two or three simultaneous speech signals, a SNR of the target speech is less than 0 dB. Though preprocessing improves the SNR of the target speech, the leak from non-target speech signals remains. In some frequency bands, the power of the leak is larger than that of the target speech, which is one of the biggest reasons for

speech recognition error. In preprocessing, white noise addition improves the SNR of such frequency bands. In addition, by masking the frequency bands, recognition performance is expected to improve.

2.1. Preprocessing for Automatic Speech Recognition

To improves the SNR of the input speech signals before performing ASR, we selected *Geometric Source Separation (GSS)* from a lot of methods to improve SNR [6, 7, 8]. GSS relaxes the limitation on the relationship between the number of sound sources and microphones. It can separate up to $N-1$ sound sources with N microphones, by introducing "geometric constraints" obtained from the locations of sound sources and the microphones. This means that GSS requires sound source directions as prior information. Given accurate sound source directions, GSS shows comparable performance with ICA. The GSS that we used was described in detail in [9]. For accurate sound source localization for GSS, we use Multiple SIgnal Classification (MUSIC)[6]. Usually multi-channel sound source separation techniques such as GSS cause spectral distortion. Such a distortion affects acoustic feature extraction for ASR, especially the normalization processes of an acoustic feature vector, because the distortion causes fragmentation of the target speech in the spectro-temporal space, and produces a lot of sound fragments. To reduce the influence of spectral distortion for ASR, we employed two techniques; a multi-channel post-filter and white noise addition.

2.1.1. Multi-Channel Post-Filter for GSS

The multi-channel post-filter [9] is used to enhance the output of GSS. It is based on the optimal estimator originally proposed by Ephraim and Malah [10]. Their method is a kind of spectral subtraction [11], but it generates less distortion because it takes temporal and spectral continuities into account. We extend their method to enable support of multi-channel signals so that they can estimate both stationary and non-stationary noise. In other words, the noise variance estimation $\lambda_m(k,\ell)$ is expressed as follows:

$$\lambda_m(k,\ell) = \lambda_m^{stat.}(k,\ell) + \lambda_m^{leak}(k,\ell), \tag{1}$$

where $\lambda_m^{stat.}(k,\ell)$ is one of the stationary component of the noise for sound source m at time frame ℓ for frequency k, and $\lambda_m^{leak}(k,\ell)$ is the estimate of source leakage.

We compute the stationary noise estimate, $\lambda_m^{stat.}(k,\ell)$, using the Minima Controlled Recursive Average (MCRA) technique proposed by Cohen [12]. To estimate λ_m^{leak}, we assume that the interference from other sources is reduced by a factor η (typically $-10\ \mathrm{dB} \le \eta \le -5\ \mathrm{dB}$) by GSS. The leakage estimate is thus expressed as follows:

$$\lambda_m^{leak}(k,\ell) = \eta \sum_{i=0, i\neq m}^{M-1} Z_i(k,\ell), \tag{2}$$

where $Z_m(k,\ell)$ is the smoothed spectrum of the m^{th} source, $Y_m(k,\ell)$. It is recursively defined as follows:

$$Z_m(k,\ell) = \alpha_s Z_m(k,\ell-1) + (1-\alpha_s) Y_m(k,\ell). \tag{3}$$

Thus, a posteriori SNR $\gamma(k,\ell)$ is estimated as a power ratio of the input signal, and the estimated noise is denoted by

$$\gamma(k,\ell) = \frac{|Y_m(k,\ell)|^2}{\lambda_m(k,\ell)}. \tag{4}$$

A priori SNR is estimated by the following equations:

$$\begin{aligned} \xi(k,\ell) &= \alpha_p G_{H1}^2(k,\ell-1)\gamma(k,\ell-1) \\ &\quad +(1-\alpha_p)\max\{\gamma(\mathrm{k},\ell)-1, 0\} \end{aligned} \tag{5}$$

$$\alpha_p = \left(\frac{\xi(k,\ell-1)}{1+\xi(k,\ell-1)}\right)^2 + \alpha_{min}, \tag{6}$$

where $G_{H1}(\cdot)$ is the spectral gain function when speech exists defined by the following equation:

$$G_{H1}(k,\ell) = \frac{\xi(k,\ell)}{1+\xi(k,\ell)} \exp\left\{\frac{1}{2}\int_{\frac{\xi(\mathrm{k},\ell)}{1+\xi(\mathrm{k},\ell)}\gamma(\mathrm{k},\ell)}^{\infty} \frac{\mathrm{e}^{-\mathrm{t}}}{\mathrm{t}}\right\}. \tag{7}$$

Finally, the probability of speech presence is calculated as

$$\begin{aligned} p(k,\ell) &= \left\{1 + \frac{\hat{q}(k,\ell)}{1-\hat{q}(k,\ell)}(1+\xi(k,\ell))\right. \\ &\quad \left.\exp\left(-\frac{\xi(k,\ell)}{1+\xi(k,\ell)}\gamma(k,\ell)\right)\right\}^{-1}, \end{aligned} \tag{8}$$

where $\hat{q}(\cdot)$ is an a priori probability of speech absence defined in [9].

The resulting post-filter, thus, improves the SNR of speech separated by spectral subtraction based on $p(k,\ell)$. Please note that $p(k,\ell)$ is obtained by estimating two types of noises with a microphone array. Most conventional post-filters focus on the reduction of only one type of noise, i.e., stationary background noise [13].

2.1.2. White Noise Addition

Further reduction of spectral distortion caused by sound source separation is exploited by using the psychological evidence that noise helps perception, which is known as *auditory induction.* This evidence is also useful for ASR, because an additive noise plays a roll to blur the distortions, that is, to avoid the fragmentation. Actually, the addition of a colored noise has been reported to be effective for noise-robust ASR [14]. They added office background noise after spectral subtraction, and showed the feasibility of this technique in noisy speech recognition.

We exploit covering a distortion in any frequency band by adding a white noise, a kind of broad-band noises, to noise-suppressed speech signals. In accordance with this addition,

we use an acoustic model trained with clean speech and white-noise-added speech. Thus, the system is able to assume only one type of noise included in speech, that is, white noise. It is easier for ASR to deal with one type of noise than various kinds of noises, and white noise is suitable for ASR with a statistical model.

2.2. Missing-Feature-Theory (MFT) Based Integration

Several robot audition systems with preprocessing and ASR have been reported so far [15, 16]. Those systems just combined preprocessing with ASR and focused on the improvement of SNR and real-time processing. Most reports on MFT have focused on a single channel input, so far [4, 5]. It is difficult to obtain information enough to estimate the reliability of acoustic features in a single channel approach. On the other hand, McCowan *et al.* reported a technique of noise-robust ASR using a combination of microphone array processing and MFT[17]. Their target was a speech mixed with a low level of background speech. However, our target is a mixture of two or three speech signals of which the levels are the same. Therefore, we integrated preprocessing and ASR for a mixture of speech using MFT.

MFT uses *missing feature masks (MFMs)* in a temporal-frequency map to improve ASR. Each MFM specifies whether a spectral value for a frequency bin at a specific time frame is reliable or not. Unreliable acoustic features caused by errors in preprocessing are masked using MFMs, and only reliable ones are used for a likelihood calculation in the ASR decoder. The decoder is an HMM-based recognizer, which is commonly used in conventional ASR systems. The estimation process of output probability in the decoder is modified in MFT-ASR.

Let $M(i)$ be a MFM vector that represents the reliability of the i-th acoustic feature. The output probability $b_j(x)$ is given by the following equation:

$$b_j(x) = \sum_{l=1}^{L} P(l|S_j) \exp\left\{ \sum_{i=1}^{N} M(i) \log f(x(i)|l, S_j) \right\}, \tag{9}$$

where $P(\cdot)$ is a probability operator, $x(i)$ is an acoustic feature vector, N is the size of the acoustic feature vector, and S_j is the j-th state.

MFT-based methods show high robustness against both stationary and non-stationary noises when the reliability of acoustic features is estimated correctly. The main issue in applying them to ASR is how to estimate the reliability of input acoustic features correctly. Because the distortion of input acoustic features are usually unknown, the reliability of the input acoustic features cannot be estimated. To estimate MFM, we used *Mel-Scale Log Spectrum (MSLS)* [18] as an acoustic feature and developed an automatic MFM generator based on the multi-channel post-filter.

2.2.1. Design of features: Mel-Scale Log Spectrum

To estimate reliability of acoustic features, we have to exploit the fact that noises and distortions are usually concentrated in some areas in the spectro-temporal space. Most conventional ASR systems use *Mel-Frequency Cepstral Coefficient (MFCC)* as an acoustic feature, but noises and distortions are spread to all coefficients in MFCC. In general, Cepstrum based acoustic features like MFCC are not suitable for MFT-ASR, Therefore, we use *Mel-Scale Log Spectrum* (MSLS) as an acoustic feature.

MSLS is obtained by applying inverse discrete cosine transformation to MFCCs Then three normalization processes are applied to obtain noise-robust acoustic features; C0 normalization, liftering, and Cepstrum mean normalization. The spectrum should be transformed into the cepstrum once since these processes are applied in a cepstral domain.

2.2.2. Automatic MFM generator

We developed an automatic MFM generator by using GSS and a multi-channel post-filter with an 8-ch microphone array.

The missing feature mask is a matrix representing the reliability of each feature in the time-frequency plane. More specifically, this reliability is computed for each time frame and for each Mel-frequency band. This reliability can be either a continuous value from 0 to 1 (called "*soft mask*"), or a binary value of 0 or 1 (called "*hard mask*"). In this paper, hard masks were used.

We compute the missing feature mask by comparing the input and the output of the multi-channel post-filter presented in Section 2.1.1. For each Mel-frequency band, the feature is considered reliable if the ratio of the output energy over the input energy is greater than threshold T. The reason for this choice is based on the assumption that the more noise present in a certain frequency band, the lower the post-filter gain will be for that band. The continuous missing feature mask $m_k(i)$ is thus computed as follows:

$$m_k(i) = \frac{S_k^{out}(i) + N_k(i)}{S_k^{in}(i)}, \tag{10}$$

where $S_k^{in}(i)$ and $S_k^{out}(i)$ are the post-filter input and output energy for frame k at Mel-frequency band i, and $N_k(i)$ is the background noise estimate for that band. The main reason for including the noise estimate $N_k(i)$ in the numerator of Eq. (10) is that it ensures that the missing feature mask equals 1 when no speech source is present. Finally, we derive a hard mask $M_k(i)$ as follows:

$$M_k(i) = \begin{cases} 1 & \text{if } m_k(i) > T, \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where T is an appropriate threshold.

To compare our MFM generation with an ideal MFM, we use *a priori* MFMs, which is defined as follows:

$$M_k(i) = \begin{cases} 1 & \text{if } |S_k^{out}(i) - S_k(i)| < T' \\ 0 & \text{otherwise} \end{cases}, \tag{12}$$

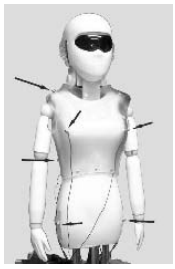


Fig. 1. SIG2 with 8 microphones

Fig. 2. ASIMO with 8 microphones

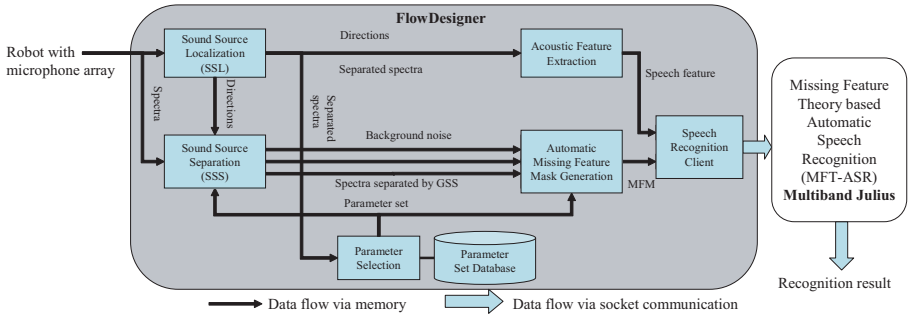


Fig. 3. Overview of the real-time robot audition system

where $S_k(i)$ is the spectrum of the clean speech that corresponds to $S_k^{out}(i)$, and T' is 0.5 in our experiments.

3. SYSTEM IMPLEMENTATION

This section explains the implementation of the real-time robot audition system. Fig. 1 and 2 show an 8-ch microphone array embedded in Humanoid SIG2 and Honda ASIMO, respectively. The positions of the microphones are bilaterally symmetric for the both robots. This is because the longer the distance between microphones is, the better the performance of GSS is. Fig. 3 depicts the architecture of the system. It consists of six modules: Sound Source Localization (SSL), Sound Source Separation (SSS), Parameter Selection, Acoustic Feature Extraction, Automatic Missing Feature Mask Generation, and Missing Feature Theory based Automatic Speech Recognition (MFT-ASR). The five modules except for MFT-ASR are implemented as component blocks of *FlowDesigner* [19], a free data flow oriented development environment.

4. EVALUATION

We evaluated the robot audition system on the following points:

1. Recognition performance of simultaneous speech,
2. Processing speed, and
3. Application to a rock-paper-scissors game using only speech information.

4.1. Recognition of Simultaneous Speech Signals

4.1.1. Evaluation of MFT and white noise addition

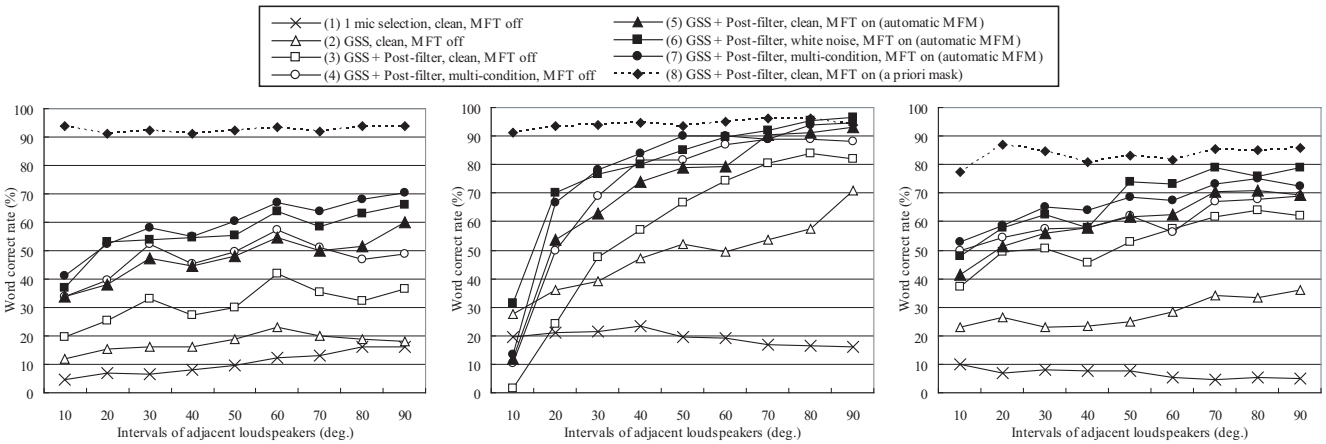
To evaluate how MFT and white noise addition improve the performance of automatic speech recognition, we conducted isolated word recognition of three simultaneous speech. In this experiment, Humanoid SIG2 with an 8-ch microphone array was used in a 4 m $\times$ 5 m room. Its reverberation time (RT_{20}) was 0.3–0.4 seconds.

Three simultaneous speech for test data were recorded with the 8-ch microphone array in the room by using three loudspeakers (Genelec 1029A). The distance between each loudspeaker and the center of the robot was 2 m. One loudspeaker was fixed to the front (center) direction of the robot. The locations of left and right loudspeakers from the center loudspeaker varied from ± 10 to ± 90 degrees at the intervals of 10 degrees. ATR phonemically-balanced word-sets were used as a speech dataset. A female (f101), a male (m101) and another male (m102) speech sources were used for the left, center and right loudspeakers, respectively. Three words for simultaneous speech were selected at random. In this recording, the power of robot was turned off.

By using the test data, the system performed isolated word recognition of three simultaneous speech signals. The size of vocabulary was 200 words. The eight conditions of the experiments are as follows:

(1) The input from the left-front microphone was used without any processing and MFT using **a clean acoustic model**.
(2) Only GSS was used as preprocessing. The **clean acoustic model** was used.
(3) GSS and Post-filter were used as preprocessing, but MFT function was not. The clean acoustic model was used.
(4) The same condition as (3) was used except for the use of a **multi-condition-trained (MCT) acoustic model**.
(5) The same condition as (3) was used except for the use of **MFT function with automatically generated MFM**.
(6) The acoustic model trained with *white-noise-added speech* (**WNA acoustic model**) was used. Except for this, the condition was the same as (5).
(7) The **MCT acoustic model** was used. The other conditions were the same as (5). This is for comparison with the **WNA acoustic model**.
(8) The same condition was used except for the use of *a priori* **MFM**.

The **clean acoustic model** was trained with 10 male and 12 female ATR phonemically-balanced word-sets excluding the three word-sets (f101, m101, and m102) which were used for the recording. Thus, it was a speaker-open and word-closed acoustic model. The **MCT acoustic model** was trained with the same ATR word-sets as mentioned above, and separated speech datasets. The latter sets were generated by separating three-word combinations of f102-m103-m104 and f102-m105-m106, which were recorded in the same way as the test data. The **WNA acoustic model** was trained with the same ATR wordsets as mentioned above, and the clean speech



(a) The left speaker (b) The center speaker (c) The right speaker

Fig. 4. Word correct rates of three simultaneous speakers with our system

Table 1. Word correct rate (WCR in %) of the center speaker according to each localization method

Acoustic model	*White noise addition*			*Clean model*		
Interval	30°	60°	90°	30°	60°	90°
given	90.0	88.5	91.0	85.0	84.5	87.0
steered BF	82.3	90.5	89.0	65.5	70.6	72.4
MUSIC	86.0	83.3	86.7	57.0	74.0	64.5

to which white noise was added by 40 dB of peak power. Each of these acoustic models was trained as 3-state and 4-mixture triphone HMM, because 4-mixture HMM had the best performance among 1, 2, 4, 8, and 16-mixture HMMs.

The results were summarized in Fig. 4. MFT-ASR with Automatic MFM Generation outperformed the normal ASR. The **MCT acoustic model** was the best for MFT-ASR, but the **WNA acoustic model** performed almost the same. Since the **WNA acoustic model** does not require prior training, it is the most appropriate acoustic model for robot audition. The performance at the interval of 10-degree was poor in particular for the center speaker, because any current sound source separation methods fails in seprating such close three speakers. The fact that *A priori* mask showed a quite high performance may suggest not a few possibilities to improve the algorithms of MFM generation.

4.1.2. Evaluation of Sound Source Localization Effects

This section evaluates how the quality of sound source localization methods including manually given localization, steered Beamformer and MUSIC affects the performance of ASR. SIG2 used steered BF. Since the performance MUSIC depends on the number of microphones on the same plane, we used Honda ASIMO shown in Fig. 2, which was installed in a 7 m × 4 m room. Its three walls were covered with sound absorbing materials, while the other wall was made of glass which makes strong echoes. The reverberation time (RT_{20}) of the room is about 0.2 seconds. We used the condition (6) in Section 4.1.1, and used three methods of sound source localization with clean and WNA acoustic models.

Table 2. Processing time (Pentium4 2.4 GHz)

input signal	800 sec	
total process time	499 sec	(realtime factor:0.62)
preprocess	369 sec	(CPU load: 50-80%)
ASR	130 sec	(CPU load: 30-40%)
output delay	0.446 sec	

The results of word correct rates were summarized in Table 1. With the **clean acoustc model**, MUSIC outperformed steered BF, while with the **WNA acoustic mode**, the both performances were comparable. In case of **given localization**, improvement by white noise addition training was small. On the other hand, training with white noise addition improved word correct rates greatly for both steered beamformer and MUSIC. The main reason to cause the poor performance is distortion in SSS. This distortion increases by two factors: SSL errors, and non-linear distortion in Post-Filter. The non-linear distortion becomes larger when the quality of sound separated by GSS is worse. In other words, it depends on the SSL errors. This means that the distortion is mainly caused by the SSL errors. Actually, in the results with **clean acoustic model**, the ASR performance with steered BF and MUSIC is 10 – 30 pts worse than that with given localization. On the other hand, **WNA acoustic model** improves the performance up to almost the same as given localization.

4.2. Processing Speed

We measured processing time when our robot audition system separated and recognized speech signals of 800 seconds shown inTab. 2. As a whole, our robot audition system ran fast as in real time.

4.3. Application to A Rock-Paper-Scissors Game

As an application of our robot audition system, we demonstrate a rock-paper-scissors game that includes a recognition task of three simultaneous utterances. The room was the same

a) U2:Let's play rock-paper-scissors. b) U1-U3: rock-paper-scissors... c) U1: paper, U2: paper, U3: scissors d) A: U3 won.

Fig. 5. Snapshots of rock-paper-scissors game (A: ASIMO, U1:left user, U2:center user, U3: right user)

as the other experiments. ASIMO was located at the center of the room, and three speakers stood 1.5 m away from ASIMO at 30 degree intervals. A speech dialog system which is specialized to this task was connected with our robot audition system. ASIMO judged who won the game by using only speech information. Note that no visual information was used in this task. Because they said rock, paper, or scissors simultaneously in a environment where robot noises exist, the SNR input sound was less than -3 dB. All of the three utterances had to be recognized successfully to complete the task.

Fig. 5 shows a sequence of snapshots for a trial of this task. In this case, a unique winner existed, but the system was able to cope with drawn cases. The system had no problem in the case of another layout of speakers as long as they did not stand in the same direction. Since the number of speakers was detected in SSL, the cases of two speakers were also supported. Theoretically, more than three speakers can be supported, but the performance becomes worse. The task success rate is not evaluated in detail. However, it is around 60% and 80% in the cases of three and two speakers, respectively.

5. CONCLUSION

We reported the robot audition system that recognizes speech that is contaminated by simultaneous speech. The system is based on two key ideas – preprocessing of ASR and missing-feature-theory based integration of preprocessing and ASR. We showed the effectiveness of the system through several experiments and a demonstration, and the conventional noise-robust ASR approaches such as only the use of a multi-condition trained acoustic model, and/or a single channel preprocessing had difficulty in achieving robot audition.

6. REFERENCES

[1] K. Nakadai *et al.*, "Active audition for humanoid," *Proc. of AAAI-2000*, 832–839.

[2] R. P. Lippmann *et al.*, "Multi-styletraining for robust isolated-word speech recognition," *Proc. of ICASSP-1987*, 705–708.

[3] C. J. Leggetter and P. C. Woodland, "Maximam likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, **9** (1995) 171–185.

[4] J. Barker *et al.*, "Robust ASR based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise," *Proc. of Eurospeech-2001*, 213–216.

[5] M. Cooke *et al.*, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Comm.*, **34**:3 (2000) 267–285.

[6] F. Asano *et al.*, "Sound source localization and signal separation for office robot "Jijo-2"," *Proc. of IEEE MFI-1999*, 243–248.

[7] C. Jutten and J. Herault, "Blind separation and sources," *Signal Processing*, **24**:1 (1995) 1–10.

[8] L. C. Parra and C. V. Alvino, "Geometric source separation: Mergin convolutive source separation with geometric beamforming," *IEEE Trans. on Speech and Audio Processing*, **10**:6 (2002) 352–362.

[9] S. Yamamoto *et al.*, "Enhanced robot speech recognition based on microphone array source separation and missing feature theory," *Proc. of IEEE ICRA-2005*, 1489–1494.

[10] Y. Ephraim and D. Malah, "Speech enhancement using minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, **ASSP-32**:6 (1984) 1109–1121.

[11] S. F. Boll, "A spectral subtraction algorithm for suppression of acoustic noise in speech," *Proc. of ICASSP-1979*, 200–203.

[12] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, **81**:2 (2001) 2403–2418.

[13] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," *Proc. of ICASSP-1988*, 2578–2581.

[14] S. Yamada *et al.*, "Unsupervided speaker adaptation based on HMM sufficient statistics in various noisy environments," *Proc. of Eurospeech-2003*, 1493–1496.

[15] I. Hara *et al.*, "Robust speech interface based on audio and video information fusion for humanoid HRP-2," *Proc. of IEEE/RSJ IROS 2004*, 2404–2410.

[16] K. Nakadai *et al.*, "Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots," *Speech Comm.*, **44**:1-4 (2004) 97–112.

[17] I. McCowan *et al.*, "Improving speech recognition performance of small microphone arrays using missing data techniques," in *Proc. of ICSLP-2002*, pp. 2181–2184.

[18] Y. Nishimura *et al.*, "Noise-robust speech recognition using multi-band spectral features," *Proc. of 148th ASA Meetings*, 1aSC7, 2004.

[19] C. Côté *et al.*, "Code Reusability Tools for Programming Mobile Robots," *Proc. of IEEE/RSJ IROS 2004*. 1820–1825.