# Speech Recognition for a Robot under its Motor Noises by Selective Application of Missing Feature Theory and MLLR

*Yoshitaka Nishimura*[*1], *Mikio Nakano*[*2], *Kazuhiro Nakadai*[*2],
*Hiroshi Tsujino*[*2] *and Mitsuru Ishizuka*[*1]

[*1]Graduate School of Information Science and Technology, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
{nisshi@mi.ci., ishizuka@}i.u-tokyo.ac.jp
[*2]Honda Research Institute Japan Co., Ltd.
8-1 Honcho, Wako-shi, Saitama 351-0114, Japan
{nakano, nakadai, tsujino}@jp.honda-ri.com

## Abstract

Automatic speech recognition (ASR) is essential for a robot to communicate with people. One of the main problems with ASR for robots is that robots inevitably generate motor noises. The noise is captured with strong power by the robot's microphones, because the noise sources are closer to the microphones than the target speech source. The signal-to-noise ratio of input speech becomes quite low (less than 0 dB). However, it is possible to estimate the noise by using information on the robot's own motions and postures, because a type of motion/gesture produces almost the same pattern of noise every time it is performed. This paper proposes a method to improve ASR under motor noises by using the information on the robot's motion/gesture. The method selectively uses three techniques – multi-condition training, maximum-likelihood linear regression (MLLR), and missing feature theory (MFT). The former two techniques cope with the motor noises by selecting the noise-type-dependent acoustic model corresponding to a performing motion/gesture. The last technique extracts unreliable acoustic features in an input sound by matching the input with a pre-recorded noise of the current motion/gesture, and masks them in speech recognition to improve ASR performance. Because, in our method, ASR technique selection affects the systems performance, we evaluated the performance of three ASRs for each noise type of a robot's motion/gesture to obtain the best technique selection rule. The preliminary results of isolated word recognition showed the effectiveness of our method using the obtained technique selection rule.

## 1. Introduction

To make human-robot communication natural, it is necessary for the robot to recognize speech even while it is moving and performing gestures. For example, a robot's gesture is considered to play a crucial role in natural human-robot communication [8, 9]. In addition, robots are expected to perform tasks by physical actions [13] to make a presentation [10]. If the robot can recognize human interruption speech while it is executing physical actions or making a presentation with gestures, it would make the robot more useful.

ASR by robot is difficult, however. This is because motor noises are inevitably generated while in motion. In addition, the power of the motor noises is stronger than that of target speech because the motors are closer to the robot's microphones. The motor noises change irregularly so we cannot obtain satisfactory performance from ASR using a conventional noise adaptation method. So far there has not been much research on speech recognition under noises of robot motion.

One of the important differences between environmental noises and robot motor noises is that a robot can estimate its motor noises because it knows what type of motion and gesture it is performing. Each kind of robot motion or gesture produces almost the same noises every time it is performed. By recording the motion and gesture noises in advance, the noises are easily estimated.

By using this, we introduce a new method for ASR under robot motor noise. Our method is based on three techniques, namely, *multi-condition training*, *maximum-likelihood linear regression* (MLLR) [5], and *missing feature theory* (MFT) [7]. These methods can utilize pre-recorded noises as described later.

Since each of these techniques has advantages and disadvantages, whether it is effective depends on the types of motion and gesture. Thus, just combining these three techniques would not be effective for speech recognition under noises of all types of motion and gestures. We therefore propose to selectively use those methods according to the types of motion and motor noises. The result of an experiment of isolated word recognition under a variety of motion and gesture noises suggested the effectiveness of this approach.

In what follows, Section 2 discusses which of the existing noise-robust ASR techniques would be effective for robot motor noises, and Section 3 explains our method for coping with robot motor noises, with details of how we apply MFT applied using pre-recorded noises. Section 4 describes the isolated word recognition experiments, and Section 5 discusses the results, before summarizing and mentioning future work in Section 6.

## 2. Noise-robust automatic speech recognition

So far, a lot of noise-robust ASR techniques have been proposed. This section discusses which techniques are suitable for ASR under robot motor noises.

A common technique is *multi-condition* training. It trains the

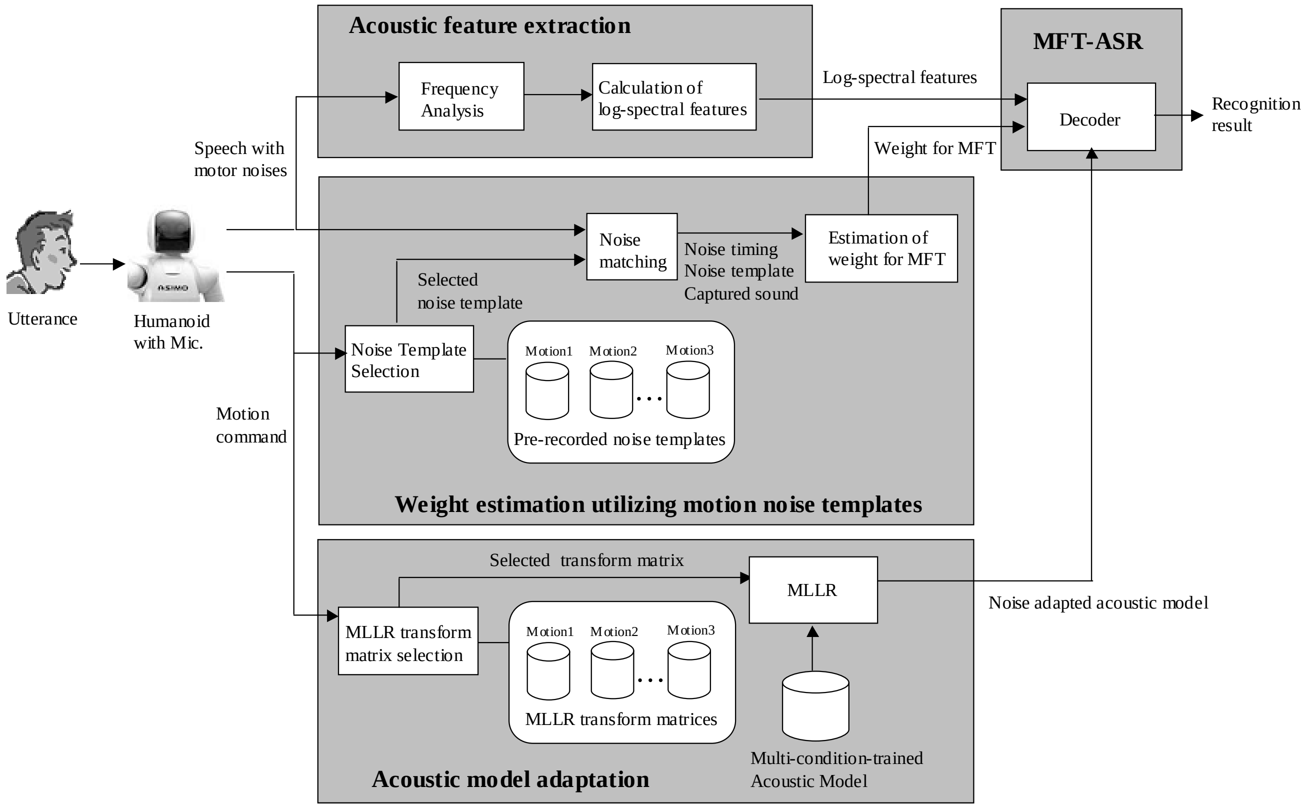
Acoustic feature extraction
Frequency Analysis
Calculation of log-spectral features
Log-spectral features
MFT-ASR
Decoder
Recognition result
Weight for MFT
Speech with motor noises
Noise matching
Noise timing
Noise template
Captured sound
Estimation of weight for MFT
Selected noise template
Utterance
Humanoid with Mic.
Noise Template Selection
Motion1
Motion2
Motion3
Pre-recorded noise templates
Motion command
Weight estimation utilizing motion noise templates
Selected transform matrix
MLLR
Noise adapted acoustic model
MLLR transform matrix selection
MLLR transform matrices
Multi-condition-trained Acoustic Model
Acoustic model adaptation

Figure 1: *Block diagram of the proposed method*

acoustic model on speech data to which noises are added. This technique improves ASR performance when an input signal includes the noises added in training the acoustic model. This has a characteristic that it is easy to cope with stationary noises rather than non-stationary ones. So, we expect that this is effective for speech recognition in performing a motion or a gesture that produces stationary noises.

MLLR also improves the robustness of ASR by using an adaptation technique with the affine transform. MLLR adaptation for a multi-condition acoustic model is more effective in speech recognition than that for an acoustic model trained on clean speech, because the performance of speech recognition using the multi-condition acoustic model is originally higher. Actually, we confirmed this through a preliminary experiment. Preparing multi-condition acoustic models for all kinds of motor noises without using MLLR would be time-consuming. In addition, it might suffer from overfitting.

*Missing Feature Theory (MFT)* [7] is proposed to cope with noisy speech input. When there are noises, some areas in the spectro-temporal space of speech are unreliable as acoustic features. Ignoring reliable areas or estimating features in the unreliable parts using reliable areas make it possible to perform noise-robust speech recognition. As a similar approach, multi-band ASR [11, 12] has been proposed. This method uses HMMs for each sub-band, and obtains integrated likelihood by assigning smaller weights to unreliable sub-bands. In this paper, when we use the term MFT, it includes the multi-band ASR method.

MFT-based methods show high noise-robustness against both stationary and non-stationary noises when the reliability of acoustic features is estimated correctly. One of the main issues in applying them to ASR is how to estimate the reliability of input acoustic features correctly. Because the *signal-to-noise ratio (SNR)* and the distortion of input acoustic features are usually unknown, the reliability of the input acoustic features cannot be estimated. However, because pre-recorded noises are available in recognition, the reliability estimation of the input acoustic features is easier even when the noise power is high. So, we think that MFT is more suitable to deal with the non-stationary noises from the robot's motors.

*Spectral Subtraction (SS)* [6] is one of the common methods to suppress noises. Ito *et al.* proposed to apply SS to cope with the robot's own motor noise [1]. Their method estimated the motor noise from the robot's joint angles with a neural network, and performed SS using the estimated noise. One problem with this approach is that ASR performance degraded when the noise is not well-estimated. In addition, when the noise estimation fails, the degradation is worse than that in the case of MFT approaches, because SS modifies acoustic feature directly. Since the same types of motions do not always generate the exactly-identical motor noises, it is difficult to estimate the motor noises well enough for SS to cope with noises properly. So, the SS-based method is not suitable for the robot.

When multiple microphones are available, it is possible to use speech separation techniques to extract the target speech such as *Beam Forming (BF)* [14], *Independent Component Analysis (ICA)*

[15], and *Geometric Source Separation (GSS)* [2]. BF is a common method to separate sound sources by using multiple microphones. However, in the cases of conventional BF approaches, separate speech is distorted by noises and inter-channel leak energy. This degrades ASR performance. Some BF methods with less distortion such as adaptive beamforming require a lot of computational power, which makes real-time sound source separation difficult. ICA is one of the best methods for sound source separation. It assumes that sound sources are mutually independent and the number of sound sources equals to that of microphones. These assumptions are, however, too strong to separate sound sources in the real world. In addition, it has some other problems called permutation problem and scaling problem that are hard to solve. In GSS, the limitation of the relationship between the number of sound sources and microphones is relaxed. It can separate up to $N-1$ sound sources where $N$ is the number of microphones by introducing "geometric constraints" obtained from the locations of sound sources and microphones. Actually, Yamamoto *et al.* reported a robot audition system that recognized simultaneous speech by combining of GSS and MFT-based ASR [2]. They showed the effectiveness of GSS as well as MFT-based ASR with automatic reliability estimation using the inter-channel leakage energy. However, in GSS, errors in geometric constraints affect the performance badly, while microphone and sound source locations generally include some errors in measurement and localization.

Multi-channel approaches are effective when sound source separation works properly. However, every approach generates separation errors more or less. In addition, the size of a total system tends to be large. This means that the number of parameters for the system increases and more computational power is required by the system. Because the room and computational power a robot can use are limited, they are hard problems when being applied to a robot. Therefore, we focus on single channel approaches in this paper.

Consequently, we use multi-condition acoustic model training, MLLR, and MFT. The details of our utilization of these techniques are described in the next section.

# 3. Automatic speech recognition based on missing feature theory for motor noises

## 3.1. Selective application of noise-robust ASR techniques

This section describes the proposed method using multi-condition acoustic model training, MLLR, and MFT to cope with noises generated by a robot's motion. Figure 1 illustrates the block diagram of the proposed method.

As acoustic features, we use log-spectral features, not mel-frequency cepstrum coefficient (MFCC). This is because log-spectral features are suitable for MFT as explained later. The acoustic model is trained on the speech to which noises of all kinds of motions and gestures are added.

For each type of motion, an MLLR transformation matrix for the multi-condition acoustic model is learned using some amount of speech data. When recognizing speech contaminated by a motor noise, the MLLR transformation matrix for the corresponding motion type is applied.

In addition, the pre-recorded noise for the motion is selected from pre-recorded noise templates. The pre-recorded noise is matched to the target sound which is a mixture of speech and motor noise, and which frequency band of which time frame is damaged by the motor noise for determining weights for MFT. The details of this process is described later.

As discussed in the previous section, these three techniques have advantages and disadvantages. Multi-condition training would be effective for all noises, but it might not be sufficient to adapt to each noise. MLLR enables adaption to each kind of noise, but, since MLLR's transform stays the same for all intervals of each speech, it might not work well for noises that change irregularly. MFT is expected to work well for such irregular noises, but if the difference between pre-recorded noise and the noise included in the target speech is big, MFT is not effective.

We therefore suspect that each of these are suitable for some types of noises and not suitable for other noises. We apply these techniques *selectively* according to the types of noises (Figure 2). When the robot is performing a motion or a gesture and one of the techniques has been found to be effective for the noise of that motion/gesture, that technique is applied. By this selective application, we can avoid ASR performance degradation caused by applying techniques that are not suitable for the noise.

## 3.2. Missing feature theory for motor noises

Here we describe in detail how we apply MFT by using pre-recorded noises.

As stated earlier, throughout our method, we use log-spectral features as acoustic features [3, 4]. The reason for this is as follows. Motor noise to cope with are additive noises. To use the MFT for additive noises directly, we use log-spectrum acoustic feature vectors. A log-spectral acoustic feature vector is normalized in the log-spectrum domain while MFCCs are normalized in the cepstrum domain. The performance of ASR with the log-spectral acoustic feature vector is equivalent to that with MFCC shown in Section 4. So, we use the log-spectral acoustic feature vectors.

In MFT, reliable features of the acoustic feature vector have large weight values and unreliable features have small weights. The weights affect the acoustic likelihood as described in [3, 4]. When not using MFT, the acoustic likelihood of a phoneme model $q_k$ and the acoustic feature vector $\mathbf{s_t}$ is defined by

$$L(\mathbf{s_t}|q_k) = \sum_{i=1}^{N} L(s_{ti}|q_k). \tag{1}$$

In MFT, using a weight $\omega_i$, the acoustic likelihood is defined by

$$L(\mathbf{s_t}|q_k) = \sum_{i=1}^{N} \omega_i L(s_{ti}|q_k). \tag{2}$$

Weights for MFT are determined based on the noise level. Let the log-spectrum of the estimated noise be $n(f,t)$, where $f$ is the feature index in the log-spectrum acoustic feature vector, and $t$ is the time frame. Because the range of log-spectrum is wide, we use the sigmoid function to limit the range of log-spectrum from 0 to 1. The average noise power at each frame is subtracted from the acoustic feature vector in order not to bias the value of output from the sigmoid function.

$F$ is the number of dimensions of acoustic feature vector.

$$n'(f,t) = n(f,t) - \frac{1}{F}\sum_{g=1}^{F} n(g,t) \tag{3}$$

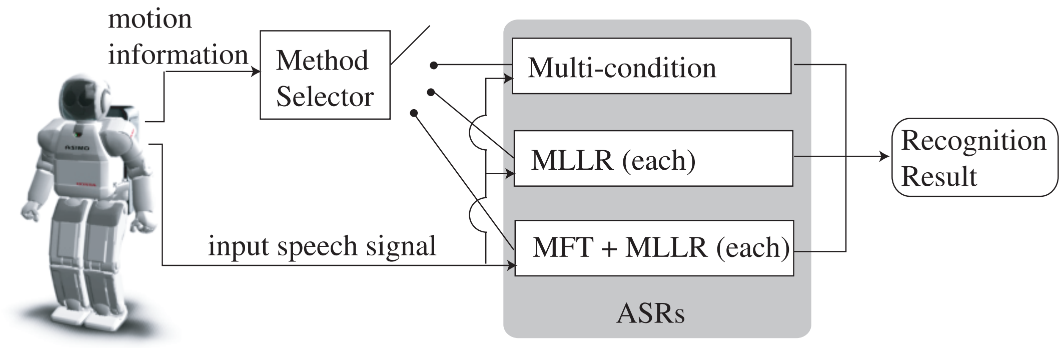

Figure 2: *Robust ASR technique selection according to the types of noise*

Next, $n'(f,t)$ is inputted to the sigmoid function. The reliability is defined by

$$\omega(f,t) = 1 + \frac{\alpha}{1 + \exp(n'(f,t))} \qquad (4)$$

where $\alpha$ is a parameter to represent the sharpness of the reliability function $\omega$. When the $\alpha$ is large, the difference between the acoustic feature vectors becomes large, and *vice versa*. $\omega$ is normalized so that the sum of the weights at a frame can be equal to the number of dimensions described in [3, 4]. This normalization suppresses the change in optimized values of parameters such as insertion penalty. The normalized $\omega$ is used for MFT.

When we use a multi-condition acoustic model, the stationary noises are incorporated into the acoustic model. We therefore apply MFT only when the estimated noise is stronger than an experimentally-defined threshold $H$.

When the types of motions are the same, the corresponding motor noises have similar spectral features. We recorded the noises of all motions beforehand. These noises are used as noise templates. We used the following method to match the noise templates and the target noises. Note that the noises contained in the target sound (a mixture of speech and noises) are called the target noises in this paper. The $N$ sample average of the difference between the noise template and the target noise $D(s)$ is defined by

$$D(s) = \frac{1}{N}\sum_{n=1}^{N} |\mathbf{T}(s)_n - \mathbf{R}_n|. \qquad (5)$$

where $\mathbf{T}$ and $\mathbf{R}$ are a noise template, and a target noise, respectively. $\mathbf{T}(s)$ or $\mathbf{T}(-s)$ means the acoustic feature vector shifted forward or backward at $s$ samples. $\mathbf{R}$ is obtained as an acoustic signal including no speech data. $\mathbf{R}$ is extracted manually in this paper.

The matched $s_m$ is defined by

$$s_m = \underset{s}{\operatorname{argmin}}\, D(s). \qquad (6)$$

The acoustic features of $\mathbf{T}(s_m)$ are sent to MFT weight calculation as $n(f,t)$ in Eq. (3) with time shift information $s_m$.

## 4. Experiment

We conducted an experiment to investigate the effectiveness of the proposed method. We used the Honda Humanoid Robot ASIMO. ASIMO has two microphones mounted on its head. We made evaluations using the data recorded from the left microphone.

The data were recorded in an anechoic room. This is because we wanted to avoid the effect of room reverbaration and other environmental noise sources so that we can verify the efficacy of our proposed method, that is, to cope with the additive motor noises.

The data contain the speech signal recorded in the condition where distance from the speech source to the microphone is with switching off ASIMO's motors. We used the ATR 216 phonetically-balanced word set and conducted isolated word recognition experiments. There are 25 speaker's data in an ATR 216 phonetically-balanced word set and 1 speaker's data consist of 216 Japanese word utterances. The duration of 1 word utterance is about 1.5 to 2 second. The speech data contains speeches of 25 speakers (12 males and 13 females). The acoustic model was trained on the data of 22 speakers, (10 males and 12 females). The unsupervised MLLR is applied to adapt to noises. The test set consists of speeches of 3 speakers (2 males and 1 female). This set is different from the training set. The noise data contain 34 kinds of noises: motor noise when ASIMO is not moving, gesture noises, noises when ASIMO is walking, and others. The SNR of each condition and motion pattern is shown in Table 2. The multi-condition acoustic model is trained on speech data to which 34 kinds of noises are added. We also used these 34 kinds of noises for the recognition experiment. The noises of these motions were recorded several times, and the noises for evaluation, multi-condition acoustic model training and template for matching are mutually exclusive.

We compared the speech recognition performances in the six conditions shown in Table 1. Since acoustic models with multi-condition training had been found effective by our preliminary experiment, we used them for all conditions. MLLR (all) means supervised MLLR for the noises of all 34 types of motions, and MLLR (each) means supervised MLLR for the noise of each type of motion. In condition C, the weights for MFT in this condition were determined by the average of the noise over time; that is, the weights were the same for all time frames. On the contrary, in condition F, the weights were computed for each time frame using the estimated noise. We also tested SS for reference. In SS, noises are estimated by the same matching algorithm as used for MFT. Since the application of MFT without MLLR resulted in worse performance than other conditions, we do not show the result of those conditions.

Table 1: *Experimental Conditions*

| Condition | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| Multi-condition | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| MLLR (all) | | ✓ | ✓ | | | |
| MLLR (each) | | | | ✓ | ✓ | ✓ |
| MFT | | | ✓ | | | ✓ |
| SS | | | | | ✓ | |

Table 2: *Signal-to-noise ratio and word accuracy*

| Motion Pattern | | SNR(dB) | Word Accuracy (%) | | | | | | Best |
|---|---|---|---|---|---|---|---|---|---|
| | | | A | B | C | D | E | F | method |
| Motor noise w/o motion | | 8.93 | 77.93 | 77.01 | 76.23 | **81.02** | 80.25 | 80.09 | D* |
| Gesture | Right hand (1) | 6.06 | 77.31 | 77.47 | 74.85 | **77.93** | 69.60 | 75.77 | D |
| | Right hand (2) | 5.13 | 74.54 | 72.53 | 73.61 | 73.46 | 72.22 | **75.31** | F |
| | Right hand (3) | 6.76 | 77.78 | 77.47 | 76.85 | 77.78 | **77.93** | 77.62 | A or D |
| | Right hand (4) | 6.99 | 77.93 | 76.85 | 75.93 | **78.40** | 75.62 | 77.62 | D |
| | Right hand (5) | 6.96 | 77.93 | 77.01 | 78.55 | 77.47 | 73.61 | **79.32** | F |
| | Left hand (1) | 6.58 | **75.31** | 74.38 | 73.92 | 75.00 | 68.67 | **75.31** | A or F |
| | Left hand (2) | 6.16 | **73.46** | 72.99 | 72.69 | 73.15 | 70.22 | 72.99 | A |
| | Left hand (3) | 6.90 | 76.85 | 76.39 | 77.62 | 77.93 | 77.16 | **79.32** | F* |
| | Left hand (4) | 6.39 | 77.31 | 76.08 | 75.00 | 76.85 | 76.08 | **78.86** | F |
| | Left hand (5) | 7.11 | **78.09** | 77.31 | 75.46 | 77.93 | 72.38 | 76.70 | A |
| | Both hands (1) | 4.31 | 70.83 | 70.52 | 70.06 | 72.07 | 66.51 | **72.99** | F |
| | Both hands (2) | 5.31 | **71.30** | 70.52 | 68.83 | 71.14 | 67.13 | 69.60 | A |
| | Both hands (3) | 5.09 | 71.60 | 69.75 | 69.91 | 71.30 | 68.67 | **71.91** | F |
| | Both hands (4) | 5.54 | 72.38 | 70.83 | 72.53 | 72.84 | 70.22 | **73.92** | F |
| | Both hands (5) | 6.39 | 75.00 | 74.54 | 73.15 | **75.46** | 71.14 | 75.31 | D |
| | Head (1) | 7.01 | **77.62** | 76.23 | 70.22 | **77.62** | 74.07 | 73.30 | A or D |
| | Head (2) | 7.39 | 74.07 | 73.15 | 69.60 | **75.15** | 74.85 | 72.99 | D |
| | Head (3) | 7.54 | 75.15 | 73.77 | 73.92 | 75.62 | 75.77 | **76.85** | F |
| | Head (4) | -0.13 | 66.82 | 65.43 | 64.51 | **68.36** | 65.74 | 67.13 | D |
| | Head (5) | -0.42 | 66.05 | 64.66 | 65.12 | 66.67 | 63.58 | **67.28** | F |
| | Head and hands (1) | 2.45 | **65.74** | 65.12 | 63.27 | 64.97 | 62.81 | 64.51 | A |
| | Head and hands (2) | 3.11 | **66.51** | 64.97 | 63.12 | 66.20 | 60.34 | 63.89 | A |
| | Head and hands (3) | 6.33 | 74.54 | 73.77 | 74.07 | 75.15 | 72.38 | **76.39** | F |
| | Head and hands (4) | 4.76 | **73.15** | 71.91 | 71.76 | 70.99 | 70.06 | 72.84 | A |
| | Bow | 7.12 | 73.30 | 73.77 | 69.75 | **75.15** | 69.44 | 70.52 | D* |
| Walking | Pattern (1) | -5.81 | 60.65 | 58.80 | 62.35 | 61.11 | 61.73 | **63.43** | F |
| | Pattern (2) | -7.06 | **59.88** | 59.26 | 55.86 | 59.41 | 52.93 | 57.87 | A |
| | Pattern (3) | -4.24 | 67.75 | 65.90 | 64.97 | **68.36** | 63.43 | 65.90 | D |
| | Pattern (4) | -4.23 | **70.37** | 68.98 | 67.13 | 68.83 | 64.51 | 69.14 | A |
| | Pattern (5) | -4.16 | 66.51 | 65.59 | 64.81 | 66.98 | 58.80 | **67.13** | F |
| | Pattern (6) | -4.85 | **66.82** | 64.66 | 63.43 | 66.51 | 58.33 | 64.51 | A |
| | Pattern (7) | -3.77 | **70.37** | 68.98 | 67.13 | 68.83 | 64.51 | 69.14 | A |
| | Pattern (8) | -4.11 | 65.90 | 64.81 | 64.81 | **66.67** | 60.49 | 65.59 | D |

* shows the best method is better than A with the significance level $p < 0.05$.

Table2 shows the experimental results. Conditions A, D, and F give better performance. In addition to multi-condition training, MLLR(each) and MFT are found effective for certain kinds of noises. On the contrary, MLLR(all) and SS are found to be not effective.

## 5. Discussion

Based on the experimental results, we can consider it possible to improve speech recognition performance by selecting condition A, D, or F according to the types of motion/gesture. This selective application of noise-robust techniques would perform better than employing a fixed strategy, that is, using one of the conditions of

A, D, and F for all types of noises.

Although applying MLLR to each noise type and applying MFT may seem effective for certain kinds of noises, the improvement is rather small. We suspect that this is because the acoustic model based on multi-condition training is already well adapted to most of the noise types. The noises which were used in multi-condition training and the noises added to the target speech were recorded in the exact same environment. This is not the case, however, in robot speech recognition in a real environment where there is reverberation and the distance between the human speaker and the robot changes. We think if the environment is different, acoustic models obtained by multi-condition training is less effective and MLLR and MFT would achieve a more statistically signigicant improvement in ASR performance.

## 6. Summary and future work

In this paper, we have proposed an automatic speech recognition method that copes with a robot's own motor noises. In order to improve ASR under a robots' own motor noises, our method used three techniques, that is, multi-condition training, MLLR adaptation, and the missing feature theory. In applying the missing feature theory, automatic estimation of unreliable acoustic features is a main issue. Our method solved this problem by utilizing information on a motion pattern obtained from a robot controller and a pre-recorded motor noise corresponding to the motion pattern. Also, it has another new feature that it selectively applies those three noise-robust techniques to according to the types of noises. The results of a preliminary experiments suggested that this method is effective.

For further improvement in ASR for a robot with motor noises, we still need to solve several problems. We should confirm the effectiveness of our method in a real environment with reverberation and in a dynamically-changing environment as mentioned in Sec. 5. In addition, it is required to improve noise estimation for the better weighting in MFT. We are also considering combining our method with sound source separation by using multi-channel microphones embedded in the robot.

## 7. Acknowledgements

The authors thank Prof. Sadaoki Furui and Dr. Koji Iwano of Tokyo Institute of Technology for their advice on MFT, and Sunichi Yamamoto of Kyoto University for his help.

## 8. References

[1] A. Ito, T. Kanayama, M. Suzuki and S. Makino, "Internal noise suppression for speech recognition by small robots," *Proc. European Conference on Speech Communication and Technology (Eurospeech-2005)*, 2005, pp. 2685–2688.

[2] S. Yamamoto, K. Nakadai, J. Valin, J. Rouat, F. Michaud, T. Ogata, K. Komatani, H. G. Okuno, "Making A Robot Recognize Three Simultaneous Sentences in Real-Time," *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2005)*, 2005, pp. 897–892.

[3] Y. Nishimura, T. Shinozaki, K. Iwano and S. Furui, "Noise-robust speech recognition using multi-band spectral features," *Proc. of 148th Acoustical Society of America Meetings*, 2004, 1aSC7.

[4] Y. Nishimura, T. Shinozaki, K. Iwano and S. Furui, "Noise-robust speech recognition using band-dependent weighted likelihood," Technical report of IEICE, 2003, SP2003-116, pp. 19–24 (*in Japanese*).

[5] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, 1995, vol.9, pp. 171–185.

[6] S. F. Boll, "Supression of acoustic noise in speech using spectral subtraction," IEEE Trans. on Acoust., Speech, Signal Process., ASSP-33, 1979, vol.27, pp. 113–120.

[7] 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," *Proc. EUROSPEECH 2001*, 2001, vol.1, pp. 213–216.

[8] C. Breazeal, C., "Designing Sociable Robots," MIT press, 2002.

[9] H. Miwa, T. Okuchi, K. Itoh, H. Takanobu and A. Takanishi, "A New Mental Model for Humanoid Robots for HumanFriendly Communication-Introduction of Learning System, Mood Vector and Second Order Equations of Emotion," *Proc. IEEE International Conference on Robotics and Automation*, 2003, pp. 3588–3593.

[10] Y. Nishimura, K. Kushida, H. Dohi, M. Ishizuka, J. Takeuchi and H. Tsujino, "Development and Psychological Evaluation of Multimodal Presentation Markup Language for Humanoid Robots," *Proc. 5th IEEE-RAS Int'l Conf. on Humanoid Robots (Humanoids-2005)*, 2005, pp. 393–398.

[11] A. Hagen and A. Morris, "Comparison of HMM experts with MLP experts in the full combination multi-band approach to robust ASR," *Proc. International Conference on Spoken Language Processing (ICSLP-2000)*, 2000, vol.1, pp. 345–348.

[12] H. Bourlard et al., "A new ASR approach based on independent processing and recombination of partial frequency bands," *Proc. International Conference on Spoken Language Processing (ICSLP-1996)*, 1996, pp. 426–429.

[13] M. Nakano, Y. Hasegawa, K. Nakadai, T. Nakamura, J. Takeuchi, T. Torii, H. Tsujino, N. Kanda, and H. G. Okuno, "A two-layer model for behavior and dialogue planning in conversational service robots," *Proc. 2005 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*, 2005, pp. 1542–1547.

[14] I. Hara, F. Asano, H. Asoh, J. Ogata, N. Ichimura, Y. Kawai, F. Kanehiro, H. Hirukawa, and K. Yamamoto, "Robust Speech Interface Based on Audio and Video Information Fusion for Humanoid HRP-2," *Proc. 2004 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS-2004)*, 2004, pp. 2404–2410.

[15] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind Source Separation Combining Independent Component Analysis and Beamforming," EURASIP Journal on Applied Signal Processing, 2003, vol. 2003, no. 11, pp. 1135–1146.