Speech Recognition for a Humanoid with Motor Noise Utilizing Missing Feature Theory

Yoshitaka Nishimura and Mitsuru Ishizuka
Graduate School of Information Science and Technology,
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, JAPAN
Email: nisshi@mi.ci.i.u-tokyo.ac.jp, ishizuka@i.u-tokyo.ac.jp

Kazuhiro Nakadai, Mikio Nakano and Hiroshi Tsujino
HONDA Research Institute Japan Co., Ltd.
8-1 Honcho, Wako-shi, Saitama 351-0114, JAPAN
Email: {nakadai,nakano,tsujino}@jp.honda-ri.com

Abstract— **Automatic speech recognition (ASR) is essential for a human-humanoid communication. One of the main problems with ASR is that a humanoid inevitably generates motor noises. These noises are easily captured by the humanoid's microphones because the noise sources are closer to the microphones than the target speech source. Thus, the signal-to-noise ratio (SNR) of input speech becomes quite low (sometimes less than 0 dB). However, it is possible to estimate these noises by using information about the humanoid's own motions and gestures. In this paper we propose a method to improve ASR for a humanoid with motor noises by utilizing the information about the humanoid's motions/gestures. The method consists of psychologically-inspired noise suppression and missing-feature-theory-based ASR (MFT-ASR). The proposed noise suppression technique adds white noise after noise suppression which does not improve SNR, but it is suitable for MFT-ASR. This is inspired by the fact that noise addition sometimes helps human perception as described in Gestalt psychology. MFT-ASR improves ASR by masking unreliable acoustic features in the input sound. The information obtained on motion/gesture is used for estimating reliability of acoustic features in MFT-ASR. We evaluated the proposed method with noisy speech recorded by Honda ASIMO in a room with reverberation. The noise data contained 32 kinds of noises: motor noises without motions, gesture noises, walking noises, and so on. The experimental results show that the proposed method outperforms the conventional multi-condition training technique.**

I. Introduction

In the future humanoids are expeted to be partners with humans. To facilitate this partnership the humanoid should be able to listen to the user's speech by using its own microphones. It is not realistic to assume that the user always wears a headset. As we develop such a humanoid, "noise" generated by its actuators is a real problem. The humanoid is basically a highly redundant system, so it includes a lot of motors as well as cooling fans for humanoid-embedded processors required to achieve human-like behaviors autonomously. These human-like behaviors are effective in making rich human-humanoid interactions. For example, a humanoid's gesture is considered to play a crucial role in natural human-humanoid communication [1], [2]. It is helpful in communicating with people for the humanoid to perform tasks and make presentations [3] accompanied by physical actions [4]. These motions, however, require high torque and high power motors, and fans which are capable of high rpms to cool the powerful CPUs. This naturally leads to loud noises. Furthermore, the actuators are closer to microphones embedded in the humanoid than the target speech source. Because of the close proximity of these noises sound signals captured with the microphones have a low *signal-to-noise ratio (SNR)* which can be less than 0 dB. In addition, the motor noises are not constant, resulting in an input SNR that changes dynamically. These factors make it difficult for the humanoid to recognize human speech while in motion. Most researchers working on human-humanoid communication tend to avoid this problem by wearing a headset to input a voice command instead of using the humanoid's own microphones [1]. Some researchers are trying to use humanoid-embedded microphones for speech recognition [5], [6]. However, they deal with stationary noises, that is, they assume that the humanoid is stationary with respect to speech recognition. One of the important differences between environmental noises and humanoid motor noises is that the humanoid can estimate its motor noises because it knows what type of motion or gesture it is performing. Each kind of motion or gesture produces a similar noise pattern every time it is performed. So, by recording the motion and gesture noises in advance, a motor noise can be easily estimated from the information on the corresponding motion or gesture.

In this paper, we propose a new method to improve *Automatic Speech Recognition (ASR)* for a humanoid with motor noises by utilizing information about the humanoid's motion/gesture. This method consists of two stages; noise suppression suitable for ASR, and ASR based on the *Missing Feature Theory (MFT)* which improves ASR by masking unreliable acoustic features in an input sound [7]. The motion/gesture information is used for estimating reliability of acoustic features for MFT. The result of the experiment on isolated word recognition under the condition where there exist a variety of motion and gesture noises supports the effectiveness of our proposed method.

The rest of this paper is organized as follows: Section II discusses which of the existing noise-robust ASR techniques would be effective for humanoid motor noises. Section III explains our method for coping with humanoid motor noises, while detailing how we apply MFT using pre-recorded noises. Section IV describes the isolated word recognition experiment. Section V discusses our results. The last two sections present our conclusions for future work.

1-4244-0200-X/06/$20.00 © 2006 IEEE

II. Noise-robust automatic speech recognition

So far, many noise-robust ASR techniques have been proposed. Generally, they fall into three categories; noise-robust acoustic models, decoder modification, and preprocessing. This section introduces these techniques and discusses which techniques are suitable for ASR under humanoid motor noises.

A. Noise-Robust Acoustic Model

A common technique is the *multi-condition* training. It trains the 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 acoustic model. However, speech data with all kinds of motor noises are necessary to train an acoustic model. Further, it is time-consuming and might suffer from overfitting.

Maximum-Likelihood Linear Regression (MLLR) [8] also improves the robustness of ASR by using an adaptation technique with the affine transform. It is less time-consuming than multi-condition training in terms of calculation. However, the cost of data preparation is the same as with multi-condition training. A large amount of speech data with motor noises is required to cope with the many different motor noises.

B. Decoder Modification

One approach to improving noise-robustness by modifying the ASR decoder is *Missing Feature Theory (MFT)* [7]. When noises exist, some areas in the spectro-temporal space of speech are unreliable as acoustic features. In MFT, such unreliable acoustic features are masked and only reliable ones are used for likelihood calculation in the ASR decoder. So, this process requires some modifications to the ASR decoder. In a similar approach, multi-band ASR [9], [10] 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 can indicate both MFT and multi-band ASR.

MFT-based methods show high noise-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 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 noise power is high. So, we think that MFT is more suitable for dealing with the non-stationary noises from the humanoid's motors.

C. Preprocessing

Preprocessing is performed to improve the SNR of the input speech signals. There are two common approaches – single channel and multi-channel approaches.

Spectral Subtraction (SS) [11] is one of the common methods to suppress noises. Ito *et al.* proposed application of SS to cope with the humanoid's own motor noise [12]. Their method estimated the motor noise from the humanoid's joint angles with a neural network, and performed SS using this estimated noise. One problem with this approach is that ASR's performance degraded when the noise was 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 features directly. Since the same types of motions do not always generate 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 humanoid.

Other noise suppression techniques also have been reported. Ephraim and Malah reported adaptive noise suppression based on a kind of spectral subtraction [13]. This method adaptively estimates a probability of speech existence based on the spectral power of a monaural input sound. According to this probability, noises included in the input are suppressed. Generally, while spectral subtraction makes musical noises and some distortions, but the noise-suppressed signal using this method includes less musical noises and distortion, because it takes temporal and spectral continuities into account.

Nakadai *et al.* reported noise cancellation by using an internal microphone located close to the noise source[14]. However, this approach has the problem of deploying microphones for noise cancellation in the case of a humanoid, because a humanoid has many degrees of freedom that produce, a lot of noise sources, and their locations change due to gestures and walking.

When multiple microphones are available, it is possible to use speech separation techniques to extract the target speech such as *Beam Forming (BF)* [5], *Independent Component Analysis (ICA)* [15], and *Geometric Source Separation (GSS)* [6]. 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 is equal to the number of microphones. These assumptions are, however, beyond the real world capability to separate sound sources. In addition, ICA has some other problems, for example a permutation problem and a scaling problem that are hard to solve. In GSS, the limitation of the relationship between the numbers 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. Yamamoto *et al.* reported a humanoid audition system that recognized simultaneous speech by the combination of GSS and MFT-based ASR [6]. They showed the effectiveness of GSS as well as MFT-based ASR with automatic reliability estimation using the inter-channel leakage energy. However,

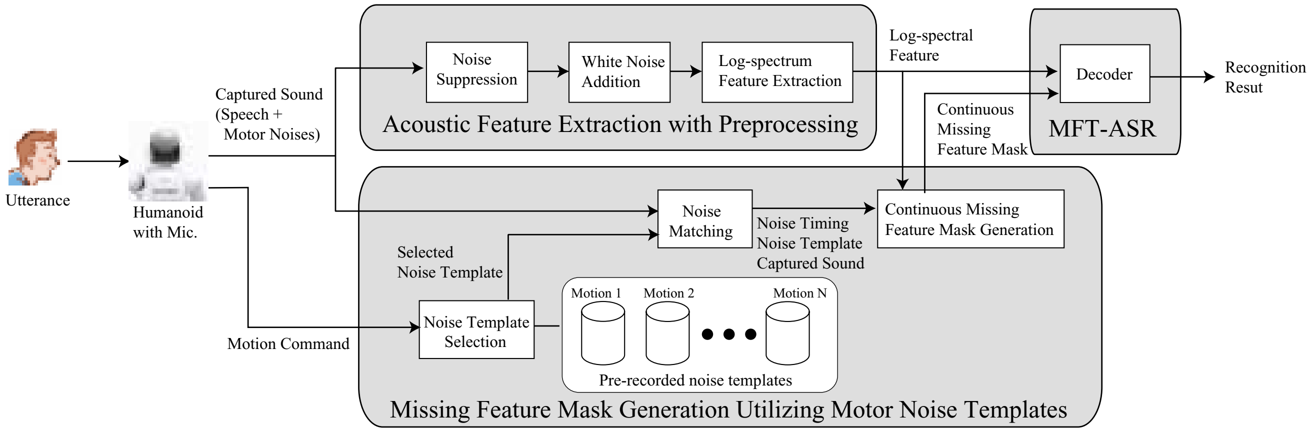
Acoustic Feature Extraction with Preprocessing
Noise Suppression
White Noise Addition
Log-spectrum Feature Extraction
Log-spectral Feature
Decoder
Recognition Resut
MFT-ASR
Continuous Missing Feature Mask
Captured Sound (Speech + Motor Noises)
Utterance
Humanoid with Mic.
Noise Matching
Noise Timing
Noise Template
Captured Sound
Continuous Missing Feature Mask Generation
Selected Noise Template
Motion Command
Noise Template Selection
Motion 1
Motion 2
Motion N
Pre-recorded noise templates
Missing Feature Mask Generation Utilizing Motor Noise Templates

Fig. 1. *Block diagram of the proposed method*

in GSS, errors in geometric constraints adversely affect the performance, while microphone and sound source locations generally include some errors in measurement and localization. Multi-channel approaches are effective when the sound source separation works properly. However, every approach more or less generates separation errors. In addition, the total system tends to be complicated. This means that the number of parameters for the system increases and more computational power is required by the system. Because the space and computational power a humanoid can provide is limited, these can be difficult problems. Therefore, in this paper we focus on single channel approaches.

Consequently, we decided to use noise suppression based on Ephraim and Malah [13] for preprocessing, and MFT [7] for decoder modification. We did not use noise-robust acoustic model training techniques such as multi-condition training and MLLR explicitly. However, the acoustic models we used in this work assume that white noise is added to speech signals. So, we trained the acoustic models on white-noise-added speech data. In this sense, we use noise-robust acoustic models.

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

Figure 1 shows the block diagram of the proposed method. It consists of three blocks – acoustic feature extraction with preprocessing, missing feature mask generation utilizing motor noise templates, and missing-feature-theory-based automatic speech recognition (MFT-ASR).

A. Acoustic Feature Extraction with Preprocessing

This block extracts acoustic features from noisy input suitable for MFT-ASR. It has three processes; noise suppression, white noise addition, and log-spectrum feature extraction.

1) Noise Suppression: The input speech has quite a low SNR of less than 0 dB. It is difficult to extract acoustic features robustly under such a noisy condition. So, first, noise suppression is performed as preprocessing of ASR. The noise suppression method we adopted is based on Ephraim and Malah's method [13] described in Sec. II.

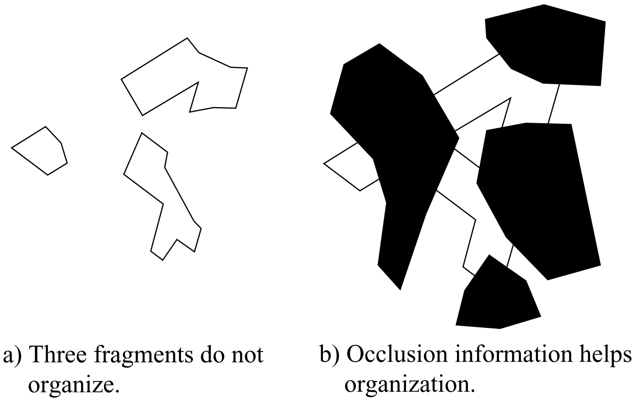
a) Three fragments do not organize.
b) Occlusion information helps organization.

Fig. 2. *An example of perceptual closure in Gestalt psychology*

2) White Noise Addition: There is no method to suppress noise without distortion. Such a distortion severely 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 produce many sound fragments.

We can learn to solve this problem from human perception mechanisms. We use the psychological evidence that noise helps perception. Figure 2 depicts an example of "perceptual closure" in Gestalt psychology [16]. Figure 2a) shows that in human perception, it is sometimes difficult to perceive organization from only fragments. Figure 2b) shows that other information such as occlusion and noise helps the organization of fragments. It is known that in the human auditory system noises that pad temporal gaps between sound fragments help auditory perception organization[16]. This is a kind of perceptual closure, and is called "auditory induction".

This evidence is also useful for ASR. We propose to add white noise to noise-suppressed speech signals. Because this process degrades speech quality in regard to SNR, one might expect that the performance of ASR would not be improved. However, it does improve the ASR performance for the following two reasons.

- An additive white noise softens the distortions. Because it is a broad-band noise, it is effective for distortion in any frequency band. Actually, we add a white noise as strong as half of the noise-suppressed signal so that the power of distortion can be ignored. Therefore, the distorted speech signal plus the white noise is regarded as non-distorted speech plus white noise.
- An acoustic model that is trained with white-noise-added speech data improves the performance of ASR for the white-noise-added speech. In this case, 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 using a statistic model.

The addition of low-level noises has been reported as an approach to noise-robust ASR in the speech community [17]. They added low-level noise to blur distortion after spectral subtraction, and showed the feasibility of this approach in noisy speech recognition. The added noise was office background noise, that is, broadband with some colors in frequency domain. So, we use this technique more aggressively to attain higher noise-robustness. The added noise power is nearly half the speech power and we use white noise instead of colored noise. As far as we know this is the first application of this techinique to a humanoid audition system. Therefore, we believe that our approach is original in this sense.

3) Log-spectrum Feature Extraction: After white noise is added, acoustic features are extracted. For acoustic features, we use log-spectral features [18], [19], not MFCC. This is because of the characteristic of motor noises. Motor noise does not have uniform power over the frequency domain. Usually the power is concentrated in certain frequency bands. This means that the effect of the motor noise depends on the frequency subband. Once it is transformed to MFCC, the motor noise spreads over all coefficients, that is, all subbands in the Cepstrum domain. The feature reliability is estimated per subband, so feature vectors in a frequency domain are suitable for MFT-ASR. In the case of MFCC, three normalization processes are performed to obtain noise-robust acoustic features; C0 normalization, liftering, and Cepstrum mean normalization. It is known that these processes are quite effective, so we conducted spectral normalization processes for log-spectral features – mean power normalization, spectrum peak emphasis and spectrum mean normalization – corresponding to the three normalization processes in MFCC. The details of spectrum normalization are described in [18], [19].

B. Missing Feature Mask Generation Utilizing Motor Noise Templates

This block estimates a missing feature mask for MFT-ASR that represents which frequency band of which time frame is damaged by the motor noise . Automatic missing feature mask generation has been studied by [20]. This estimate is still difficult without using a priori information on speech and noise. In our case, however, the system estimates motor noises by using a motion command. So, this block estimates the missing feature mask by using a motor command and pre-recorded motor noise templates. It includes three processes; noise template selection with pre-recorded motor noise templates, noise matching, and continuous missing feature mask generation.

1) Noise Template Selection: This process selects a pre-recorded noise template corresponding to an input motion command. The noise template is selected from a pre-recorded noise template database. The database is constructed by recording the noises of all motions beforehand. In our system, 32 noise templates are currently stored in the database. The selected template is sent to noise matching process.

2) Noise Matching: The inputs to this process are the selected noise template and the captured sound obtained with the humanoid's microphone. When the types of motions are the same, the corresponding motor noises have similar spectral features. So, by matching the two inputs, the target noises included in the captured sound can be estimated. Note that in this paper we call the noises contained in the target sound (a mixture of speech and noises) the target noises in this paper. We used the following method to match the noise templates and the target noises. The N sample average of the difference between the noise template and the target noise $D(s)$ is defined by

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

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.

The matched s_m is defined by

$$s_m = \operatorname*{argmin}_{s} D(s). \tag{2}$$

The acoustic features of $\mathbf{T}(s_m)$ is sent to the missing feature mask generation process with time shift information s_m.

3) Continuous Missing Feature Mask Generation: This process uses time shift information of the target noise, the selected noise template, and the captured sound, to estimate a missing feature mask for each time frame. Each value in the missing feature mask is a reliability of the corresponding subband. We can say that we use a continuous missing feature mask, because the range of the reliability is from 0 to 1.

The missing feature mask is determined based on the noise level. We define several signals here. The log-spectrum of the estimated noise $\mathbf{T}(s_m)$ is $n(k,t)$, where k is the feature index in the log-spectrum acoustic feature vector, and t is time frame. The log-spectra of the input speech and the white-noise-added signal after noise suppression are $y(k,t)$ and $p(k,t)$, respectively. The log-spectrum of the clean speech is estimated by

$$c'(k,t) = y(k,t) - n(k,t). \tag{3}$$

The weight factor $f(k,t)$ is calculated by

$$f(k,t) = \frac{|C'(k,t) - \mathrm{median}_k(C'(k,t))|}{P(k,t) - C'(k,t)} \tag{4}$$

where $\mathrm{median}_k(a(k))$ is a function that obtains the median value of $a(k)$. $P(k,t)$ and $C'(k,t)$ are normalized spectra of $p(k,t)$ and $c'(k,t)$, respectively.

Because the range of the weight factor $f(k,t)$ can be wide, we set an upper limit threshold f_{th} so that $f(k,t)$ can have a value from 0 to f_{th}. f_{th} is empirically set to 5.0. We, then, normalize it as missing feature mask $w(k,t)$, so that the sum of the $w(k,t)$ at a time frame can be equal to the number of dimensions of the acoustic feature vector K described in [18], [19]. This normalization suppresses the change in optimized values of parameters such as insertion penalty.

$$\begin{aligned} w(k,t) &= \frac{k(f,t)}{\sum\limits_{k=1}^{K} f'(k,t)} &(5)\\ f'(k,t) &= \begin{cases} f(k,t) & if \quad f(k,t) < f_{th}, \\ f_{th} & if \quad otherwise. \end{cases} \end{aligned}$$

C. Missing-Feature-Theory-Based Automatic Speech Recognition

In this block the decoder recognizes input speech based on MFT. MFT is expected to work well for irregular noises. Most distortions and noises, besides white noise, are suppressed in the first block, but the acoustic feature still includes some kind of distortion. MFT is effective in dealing with such distortions. Note that if the difference between pre-recorded noise and the noise included in the target speech is large, MFT is less effective.

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 [18], [19]. When not using MFT, the acoustic likelihood of a phoneme model q_k and the acoustic feature vector $\mathbf{s_t}$ is defined by

$$L(\mathbf{s_t}|q_k) = \sum_{i=1}^{N} L(s_{ti}|q_k). \tag{6}$$

In MFT, using a weight ω_i, the acoustic likelihood is defined by

$$L(\mathbf{s_t}|q_k) = \sum_{i=1}^{N} \omega_i L(s_{ti}|q_k). \tag{7}$$

IV. Evaluation

We evaluated the system throughout isolated word recognition to determine the effectiveness of the proposed method. We used Honda ASIMO as a testbed. ASIMO had two microphones mounted on its head. We used the data recorded through the left microphone.

We prepared two types of speech data sets for training and test data. As clean speech data, we used the ATR 216 phonetically-balanced word set. Nineteen speakers (9 males and 10 females) included in the word set were used for acoustic model training (hereafter dataset A_1). Another 3 speakers (2 males and 1 female) were used for isolated word recognition tests (hereafter dataset R_1). ASIMO has two microphones on its head, we selected ASIMO's left microphone for data capturing.

To make the training data set, we first played all speech data included in dataset A_1 through a loudspeaker, and recorded it with the left microphones in an anechoic room. The distance between ASIMO and the sound source was fixed at 100 cm, and the direction of the sound source was also fixed toward the center of ASIMO. ASIMO's stationary noise was also recorded with ASIMO on in the anechoic room. A training data set A_2 was then generated by adding the recorded speech data and noise.

The test data set was generated by performing a convolution of clean speech data and transfer functions from a sound source to ASIMO's left microphone. Motor noises were added to the convoluted speech data. The transfer functions were obtained by measurement of impulse responses. The impulse responses were measured in a 7 m (W) × 4 m (D) × 3 m (H) room. In this room, three walls of the room were covered with sound absorbing materials, and another wall was made of glass. The floor and the ceiling are flat and make echoes. There is a kitchen sink inside the room. We can hear sounds from an air-conditioner at a low frequency. So, the room has asymmetrical reverberation and a noise source in addition to the humanoid's motors. ASIMO was placed at the center of the room. The distance between ASIMO and the sound source was set at 50 cm, 100 cm, 150 cm, and 200 cm, and the direction of the sound source was fixed in direction to the front of ASIMO. The impulse response was measured at each point with ASIMO off. We also recorded 32 kinds of noises: stationary motor noise, gesture noises, and walking noises. These noise data were used not only for data set generation but also for making a pre-recorded noise template database. So, the noises of these motions were recorded several times so that the noises for test, multi-condition acoustic model training and the templates for matching would be mutually exclusive. A test data set R_2 was generated by adding the captured motor noises after convolution of R_1 and the measured transfer functions. We, thus, prepared two speech data sets: A_2 for training and R_2 for tests.

We, then, trained four triphone based acoustic models "AM-1" through "AM-4" by using the following data sets:

AM-1 the data set A_1 only (clean acoustic model),
AM-2 the data sets A_1 and A_2 (multi-condition trained acoustic model),
AM-3 the data set A_1 and a data set A_3 which was obtained by performing noise-suppression for A_2.
AM-4 the data set A_1 and a data set A_4 which was obtained by adding white noises to A_3.

Strictly, we might have to say that "AM-3" and "AM-4" are multi-condition trained models, because A_3 and A_4 still include motor noises. However, motor noises in A_3 are suppressed, so its noise level is greatly lower than A_2. A_4 is regarded as speech data with only white noise, that is, "uni-condition". So, we defined "AM-3" and "AM-4" as non multi-condition trained acoustic models.

We compared the speech recognition performances for the six conditions shown in Table I. Condition **A** is just conventional speech recognition with a clean acoustic model. In condition **B**, the system used a multi-condition trained acoustic model which is a common noise-robust technique. Most applications to robots and car navigation currently use this technique. So, we regard condition **B** as the baseline condition. In condition **C**, noise-suppressed speech signals were recognized without adding white noises by using conventional ASR. This will show the basic performance of noise suppression. In this case, we did not use mean power normalization in extracting log-spectrum acoustic features described in Sec. III-A.3, because this normalization adversely affects log-spectrum acoustic features badly due to distortions in noise suppression. Actually, we confirmed that log-spectrum acoustic features without mean power normalization outperform those with this normalization. In condition **D**, noise-suppression and white noise addition are effective, but conventional ASR was used. So, this will show the effectiveness of white noise addition. Condition **E** is the proposed method. In this condition, noise-suppression, white noise addition and MFT-ASR were performed. We expect that the performance in condition **E** to be the best among conditions **A** through **E**. Condition **F** is similar to the condition **E**. However, in missing feature mask generation, we gave the correct missing feature mask information to the system. The correct missing feature mask was generated by giving a motor noise included in the input speech as a noise template to the system. Condition **F** will exhibit the upper-limit in performance for our approach.

Table II shows the experimental results. A large bold face number denotes the best result per noise type per distance among the conditions **A** through **E**, and large italic denotes the second best result. In the columns of condition **E**, P-values[21], which denote error rates of the proposed method (condition **E**) for the baseline (condition **B**), are shown. P-values of less than 10%, which are expected to statistically improve the performance with the proposed method, were emphasized in Table II. P-values over 100% were shown as "—".

Generally, condition **F** has the best performance because it uses a priori information to estimate missing feature masks. So, when the system does not use a priori information, condition **E** is the best. Condition **B** or **D** is second best. In the cases of gestures using a hand and walking motions at 200 cm, the proposed method showed a statistically-significant improvement in ASR performance according to P-values. We could not find a significant difference in the other cases, for head gestures and walking motions at the distances of 50 cm, 100 cm, and 150 cm.

V. Discussion

The reason why the proposed method did not work well for head gestures is that head motions are not especially noisy in ASIMO, that is, for these noises the input speech has a high SNR. Actually, we could not hear the sound of head motions. This causes ASR, in the cases of these head motions, to show good performance in condition **A**. In the cases of walking motions at 50 cm, 100 cm, and 150 cm, we can also say that the proposed method did not work properly again because of high SNR input. In these cases, noise sources are a little distance away from the microphone, because the microphone was installed on the head while noises came from the legs. So, the input SNR is higher than for other gestures. However, the effect of reverberation is stronger, so condition **A** did not deal with walking motions well regardless of high SNR input. When the distance to the target speech source was 200 cm, the proposed method was more effective because input SNR was low. Thus, we can say that the proposed method is more effective than multi-condition training in the case of low SNR input, and it is comparable in the case of high SNR input.

The only use of noise suppression (condition **C**) did not produce a good performance. This means that our noise suppression method handles strong distortions well enough to affect ASR. However, the combination of noise suppression and white noise addition (condition **D**) improve ASR performance equal to multi-condition training (condition **B**). If only white noise addition is applied, the noise level is much higher than target speech signals, and speech recognition would be more difficult for the system. So, this combination use is a key technique to cope with low SNR input.

The use of MFT (condition **E**) is basically effective, especially for low SNR inputs. The results shows that the proposed method, that is, the combination of noise suppression, white noise addition and MFT is superior to multi-condition training. Compared with MFT with a priori missing feature mask (condition **F**), the proposed method is somewhat degraded by a very small amount. This means that our automatic missing feature mask generation succeeds in generating almost correct missing feature masks, and the use of pre-recorded noise templates is effective in coping with motor noises.

VI. Conclusion

In this paper, we have proposed an automatic speech recognition method that copes with a humanoid's own motor noises. In order to improve ASR when the humanoid's own motor noises are present, our method combined two techniques – noise suppression which is suitable for ASR, and missing-feature-theory-based ASR utilizing pre-recorded motor noise templates. Usually, noise suppression is a technique to improve the SNR of the input speech. For ASR, high SNR speech is not always the optimal input, because distortion by noise suppression degrades the ASR performance. We solved this problem by adding white noise to noise-suppressed signals. This idea was inspired by psychological evidence of human audio perception. 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 humanoid controller and a pre-recorded motor noise corresponding to the motion pattern.

TABLE I

Experimental Conditions

Condition	**A**	**B**	**C**	**D**	**E**	**F**
Multi-condition		✓				
Noise Suppression			✓	✓	✓	✓
White Noise Addition				✓	✓	✓
MFT					✓	
MFT (a priori mask)						✓
Acoustic Model	AM-1	AM-2	AM-3	AM-4	AM-4	AM-4

We constructed the ASR system based on the proposed method using the Honda ASIMO. The experimental results using the constructed system demonstrated that this method is effective, especially for low SNR input.

VII. Future work

For further improvement in ASR for a humanoid with motor noises, we will need to solve several problems. We should confirm the effectiveness of our method using not just recorded data but real data in a dynamically-changing environment. We are also considering combining our method with sound source separation by using multi-channel microphones embedded in the humanoid, and another new feature to select noise-robust techniques according to the types and the power levels of noises.

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.

References

[1] C. Breazeal, *Designing Sociable Robots*, MIT press, 2002.

[2] H. Miwa, T. Okuchi, K. Itoh, T. H., and A. Takanichi, "A new mental model for humanoid robots for human friendly communication – introduction of learning system, mood vector and second order equations of emotion –," in *Proc. of IEEE-RAS International Conference on Robotics and Automation (ICRA 2003)*. 2003, pp. 3588 – 3593, IEEE.

[3] 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," in *Proc. 5th IEEE-RAS International Conference on Humanoid Robots (Humanoids-2005)*, 2005, pp. 393–398.

[4] 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," in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2005)*, 2005, pp. 1542–1547.

[5] I. Hara, F. Asano, H. Asoh, J. Ogata, N. Ichimura, Y. Kawai, F. Kanehiro, H. Hirukawa, and K. Yamamoo, "Robust speech interface based on audio and video information fusion for humanoid HRP-2," in *Proc. of IEEE/RAS International Conference on Intelligent Robots and Systems (IROS-2004)*. 2004, pp. 2404–2410, IEEE.

[6] S. Yamamoto, K. Nakadai, J.-M. Valin, J. Rouat, F. Michaud, T. Ogata, H. Komatani, and H. G. Okuno, "Making a robot recognize three simultaneous sentences in real-time," in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2005)*, IEEE, Ed., 2005, pp. 897–892.

[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," in *Proc. of 7th European Conference on Speech Communication Technology (Eurospeech-2001)*. 2001, vol. 1, pp. 213–216, ESCA.

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

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

[10] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. of International Conference on Spoken Language Processing (ICSLP-1996)*, 1996, vol. 1, pp. 426–429.

[11] S. F. Boll, "A spectral subtraction algorithm for suppression of acoustic noise in speech," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP-79)*. 1979, pp. 200–203, IEEE.

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

[13] Y. Ephraim and D. Malah, "Speech enhancement using minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, 1984.

[14] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *Proc. of the Seventeenth National Conference on Artificial Intelligence (AAAI-00)*, 2000, pp. 832–839.

[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*, vol. 2003, no. 11, pp. 1135–1146, 2003.

[16] A. S. Bregman, *Auditory Scene Analysis*, The MIT Press, MA., 1990.

[17] S. Yamade, A. Lee, H. Saruwatari, and K. Shikano, "Unsupervised speaker adaptation based on hmm sufficient statistics in various noisy environments," in *Proc. of 8th European Conference on Speech Communication and Technology (Eurospeech-03)*, 2003, vol. II, pp. 1493–1496.

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

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

[20] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, 2005.

[21] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP-89)*, IEEE, Ed., 1989, pp. 532–535.

TABLE II

ISOLATED WORD RECOGNITION (% WORD CORRECT)

	50 cm							100 cm						
Condition	**A**	**B**	**C**	**D**	**E**	(P-value)	**F**	**A**	**B**	**C**	**D**	**E**	(P-value)	**F**
Motor Noise	71.45	81.02	46.45	*83.33*	**84.88**	(**0.03**)	83.18	66.20	82.25	50.00	*82.87*	**83.49**	(0.50)	83.33
Right hand (1)	59.26	74.38	24.38	*77.62*	**81.94**	(**0.00**)	82.10	44.60	68.52	20.68	*71.45*	**77.62**	(**0.00**)	79.32
Right hand (2)	53.09	71.60	16.20	*72.07*	**78.24**	(**0.00**)	78.86	39.66	*63.27*	13.73	60.34	**68.52**	(**0.01**)	73.77
Right hand (3)	65.59	80.40	30.71	**80.56**	**80.56**	(1.00)	81.94	54.78	*78.09*	28.86	75.77	**79.01**	(0.66)	78.40
Right hand (4)	61.88	79.01	33.33	*80.56*	**82.10**	(**0.10**)	85.03	50.62	72.99	29.78	*76.23*	**81.33**	(**0.00**)	81.79
Right hand (5)	62.96	76.54	24.38	**81.79**	*81.64*	(**0.01**)	82.25	48.30	72.69	19.91	*75.62*	**80.40**	(**0.00**)	79.17
Left hand (1)	68.98	79.78	37.96	*81.33*	**83.33**	(**0.05**)	83.95	59.10	*80.40*	36.27	*80.40*	**82.56**	(0.25)	81.94
Left hand (2)	70.52	81.48	45.37	*82.25*	**83.02**	(0.39)	84.41	62.04	*81.33*	43.83	78.24	**82.41**	(0.60)	82.72
Left hand (3)	67.75	79.94	37.65	*81.79*	**83.02**	(**0.08**)	83.18	54.17	*77.31*	34.72	75.31	**80.86**	(**0.07**)	79.94
Both hands (1)	56.02	74.54	24.85	*76.23*	**81.48**	(**0.00**)	80.09	40.28	66.36	21.30	*66.67*	**73.77**	(**0.00**)	75.00
Both hands (2)	60.03	75.77	26.39	*78.55*	**82.72**	(**0.00**)	83.64	47.07	69.14	25.46	*70.37*	**75.77**	(**0.00**)	76.23
Both hands (3)	53.55	72.99	26.39	*77.47*	**80.25**	(**0.00**)	79.94	45.22	66.98	20.83	*68.83*	**73.15**	(**0.00**)	73.46
Both hands (4)	59.57	74.85	27.78	*77.47*	**79.94**	(**0.01**)	81.17	47.07	70.37	24.69	*70.99*	**77.16**	(**0.00**)	79.48
Both hands (5)	59.10	74.54	22.38	*78.70*	**80.25**	(**0.00**)	81.33	46.60	69.14	18.21	*71.45*	**77.47**	(**0.00**)	79.32
Head (1)	66.82	75.62	26.39	*76.70*	**78.70**	(0.11)	83.49	56.64	**74.54**	22.84	66.05	*73.46*	(—)	79.94
Head (2)	66.82	77.47	33.80	*78.70*	**79.94**	(0.22)	83.95	58.02	**77.01**	34.72	*74.69*	*74.69*	(—)	81.02
Head (3)	70.06	*80.86*	35.80	*80.86*	**83.18**	(0.21)	82.25	64.51	**81.94**	35.65	*79.32*	78.40	(—)	82.56
Head (4)	63.12	*77.01*	30.56	75.00	**79.32**	(0.26)	79.48	48.77	*71.45*	29.78	64.35	**75.93**	(**0.03**)	76.23
Head (5)	65.12	*78.09*	28.55	76.08	**79.17**	(0.63)	81.48	56.64	*75.00*	28.86	70.22	**75.15**	(1.00)	75.31
Head and Hands (1)	67.59	*79.17*	33.80	78.70	**80.09**	(0.67)	82.72	58.33	**78.55**	32.72	74.07	*75.15*	(—)	80.56
Head and Hands (2)	60.34	74.54	22.69	*77.47*	**81.64**	(**0.00**)	81.17	44.60	66.20	21.14	*71.76*	**75.62**	(**0.00**)	76.23
Head and Hands (3)	57.25	74.54	16.67	*77.62*	**80.56**	(**0.00**)	81.02	43.67	*67.90*	14.51	66.51	**70.83**	(0.17)	75.46
Head and Hands (4)	61.11	74.23	22.22	*79.94*	**82.25**	(**0.00**)	82.41	47.69	68.36	22.53	*73.46*	**78.55**	(**0.00**)	78.40
Head and Hands (5)	62.65	78.09	30.71	*79.17*	**82.25**	(**0.03**)	83.49	50.77	72.22	27.31	*72.38*	**76.85**	(**0.02**)	79.78
Walking Motion (1)	55.25	*74.23*	25.77	71.60	**76.39**	(0.30)	79.17	44.75	**70.06**	23.61	60.03	*66.98*	(—)	73.30
Walking Motion (2)	58.95	*78.40*	28.70	70.99	**78.55**	(1.00)	78.86	47.22	**72.53**	25.46	59.57	*69.29*	(—)	72.22
Walking Motion (3)	66.51	*79.48*	27.93	78.55	**81.94**	(0.20)	81.48	53.09	**77.31**	27.47	67.90	*75.93*	(—)	77.93
Walking Motion (4)	68.83	*81.64*	38.43	81.02	**82.56**	(0.65)	82.41	56.79	*79.48*	36.73	74.54	**80.71**	(0.52)	81.17
Walking Motion (5)	64.04	*79.17*	22.84	78.70	**80.09**	(0.67)	80.71	47.22	**76.23**	20.52	66.98	*72.99*	(—)	75.00
Walking Motion (6)	63.27	77.62	23.61	*79.17*	**79.32**	(0.41)	82.07	50.00	**76.39**	20.83	68.06	*75.93*	(—)	77.16
Walking Motion (7)	68.83	*81.64*	38.43	81.02	**82.56**	(0.65)	82.41	56.79	*79.48*	36.73	74.54	**80.71**	(0.52)	81.17
Walking Motion (8)	61.27	*75.46*	22.38	75.15	**79.78**	(**0.02**)	81.02	45.37	**72.69**	19.60	64.35	*70.22*	(—)	74.69

	150 cm							200 cm						
Condition	**A**	**B**	**C**	**D**	**E**	(P-value)	**F**	**A**	**B**	**C**	**D**	**E**	(P-value)	**F**
Motor Noise	51.70	*76.70*	43.67	74.54	**78.86**	(0.26)	78.09	41.51	*69.14*	40.28	68.21	**72.22**	(0.13)	73.46
Right hand (1)	29.17	56.48	17.13	*62.04*	**69.44**	(**0.00**)	68.21	22.99	44.91	14.04	*52.47*	**60.34**	(**0.00**)	61.73
Right hand (2)	25.93	47.53	10.65	*48.61*	**57.56**	(**0.00**)	65.43	18.98	38.12	7.25	*39.51*	**50.46**	(**0.00**)	55.09
Right hand (3)	40.28	*66.82*	23.30	66.67	**71.45**	(**0.03**)	72.84	33.02	57.25	20.06	*58.95*	**65.74**	(**0.00**)	68.36
Right hand (4)	36.88	58.95	22.84	*67.28*	**72.99**	(**0.00**)	75.46	27.62	49.69	16.51	*55.09*	**65.90**	(**0.00**)	66.67
Right hand (5)	35.03	57.25	18.06	*64.66*	**73.15**	(**0.00**)	73.46	26.39	46.14	13.43	*55.86*	**63.89**	(**0.00**)	65.28
Left hand (1)	42.90	*69.75*	32.87	68.83	**73.30**	(**0.07**)	74.23	32.25	*60.19*	28.24	59.41	**67.44**	(**0.00**)	68.36
Left hand (2)	45.99	*73.30*	40.90	72.69	**75.93**	(0.20)	76.85	36.73	63.12	35.34	*63.58*	**72.22**	(**0.00**)	72.84
Left hand (3)	39.35	67.28	29.94	*68.06*	**73.61**	(**0.00**)	73.92	31.33	55.40	25.93	*58.49*	**65.28**	(**0.00**)	66.51
Both hands (1)	26.85	53.86	18.06	*56.48*	**66.36**	(**0.00**)	67.75	19.75	43.83	14.81	*45.99*	**55.40**	(**0.00**)	56.64
Both hands (2)	32.10	59.10	21.30	*60.96*	**67.28**	(**0.00**)	68.52	25.15	49.38	15.74	*51.54*	**59.10**	(**0.00**)	61.11
Both hands (3)	29.94	56.02	17.28	*58.18*	**65.28**	(**0.00**)	66.67	21.30	47.53	14.35	*48.77*	**56.64**	(**0.00**)	58.49
Both hands (4)	32.41	58.49	18.21	*61.27*	**68.06**	(**0.00**)	71.30	25.00	50.00	16.20	*51.85*	**60.49**	(**0.00**)	61.42
Both hands (5)	33.49	56.48	14.04	*61.73*	**69.44**	(**0.00**)	71.60	24.07	46.76	11.42	*51.85*	**58.02**	(**0.00**)	61.42
Head (1)	42.44	**64.81**	19.44	58.49	*62.04*	(—)	72.84	34.72	**58.64**	16.20	48.15	*57.72*	(—)	65.90
Head (2)	45.37	**67.28**	30.56	*66.05*	65.90	(—)	75.31	37.65	**61.73**	25.62	55.71	*57.72*	(—)	69.91
Head (3)	49.69	**75.31**	33.02	*70.99*	*70.99*	(—)	76.08	40.43	*64.81*	29.63	62.19	**66.67**	(0.43)	70.37
Head (4)	35.96	*60.03*	24.69	54.32	**63.43**	(0.15)	66.05	26.08	*50.00*	21.14	46.91	**57.41**	(**0.00**)	59.41
Head (5)	45.83	**66.05**	25.31	60.96	**66.05**	(—)	67.13	36.42	**58.18**	22.07	50.46	*58.02*	(—)	61.88
Head and Hands (1)	43.83	**70.37**	28.86	64.81	*67.44*	(—)	76.23	34.88	*59.57*	26.85	58.18	**61.88**	(0.30)	71.30
Head and Hands (2)	30.40	56.48	16.51	*60.80*	**65.28**	(**0.00**)	68.83	22.84	46.14	14.66	*50.15*	**58.18**	(**0.00**)	59.10
Head and Hands (3)	30.71	*56.64*	10.96	54.63	**61.73**	(**0.02**)	66.67	22.84	*46.76*	9.41	46.14	**52.93**	(**0.00**)	56.79
Head and Hands (4)	32.56	55.25	18.83	*62.04*	**69.44**	(**0.00**)	71.60	24.69	45.83	14.04	*56.94*	**63.12**	(**0.00**)	61.88
Head and Hands (5)	33.95	61.42	22.99	*64.04*	**71.30**	(**0.00**)	71.76	26.54	53.55	18.21	*56.33*	**61.57**	(**0.00**)	63.27
Walking Motion (1)	31.94	**58.18**	18.83	46.45	*57.72*	(—)	62.19	23.61	*46.60*	15.74	39.66	**51.23**	(**0.03**)	54.32
Walking Motion (2)	34.26	*62.04*	23.30	51.70	*62.65*	(0.82)	64.20	24.07	*51.54*	20.06	42.44	**53.09**	(0.50)	52.62
Walking Motion (3)	37.96	*68.52*	25.62	58.49	**70.06**	(0.50)	69.44	29.17	**58.18**	21.91	49.38	*57.87*	(—)	61.73
Walking Motion (4)	43.21	*71.45*	35.19	66.98	**73.30**	(0.39)	74.07	35.19	*61.57*	29.78	59.72	**67.75**	(**0.00**)	68.83
Walking Motion (5)	32.56	*63.89*	18.98	57.25	**64.97**	(0.65)	69.44	26.08	*51.08*	15.43	45.22	**55.86**	(**0.03**)	58.02
Walking Motion (6)	35.96	*64.51*	19.14	57.87	**65.59**	(0.65)	68.52	27.62	*55.40*	15.90	47.99	**56.02**	(0.83)	61.11
Walking Motion (7)	43.21	*71.45*	35.19	66.98	**73.30**	(0.39)	74.07	35.19	*61.57*	29.78	59.72	**67.75**	(**0.00**)	68.83
Walking Motion (8)	32.56	**60.49**	16.98	49.07	*59.57*	(—)	64.81	23.77	**49.69**	14.04	41.05	*49.38*	(—)	56.02