Whole Body Motion Noise Cancellation of a Robot for Improved Automatic Speech Recognition

Gökhan Ince, Kazuhiro Nakadai, Tobias Rodemann, Hiroshi Tsujino, Jun-ichi Imura

2011

///VSP/// Advanced Robotics 25 (2011) 1405–1426 ADVANCED ROBOTICS brill.nl/ar

Full paper

Whole Body Motion Noise Cancellation of a Robot for Improved Automatic Speech Recognition

Gökhan Ince [a,b,*], **Kazuhiro Nakadai** [a,b], **Tobias Rodemann** [c], **Hiroshi Tsujino** [a] and **Jun-ichi Imura** [b]

[a] Honda Research Institute Japan Co., Ltd, 8-1 Honcho, Wako-shi, Saitama 351-0188, Japan

[b] Department of Mechanical and Environmental Informatics, Graduate School of Information Science and Engineering, Tokyo Institute of Technology 2-12-1-W8-1, O-okayama, Meguro-ku, Tokyo 152-8552, Japan

[c] Honda Research Institute Europe GmbH, Carl-Legien Strasse 30, 63073 Offenbach, Germany

Received 16 March 2010; accepted 5 August 2010

Abstract

The motors of a robot produce ego-motion noise that degrades the quality of recorded sounds. This paper describes an architecture that enhances the capability of a robot to perform automatic speech recognition (ASR) even as the entire body of the robot moves. The architecture consists of three blocks: (i) a multi-channel noise reduction block, consisting of microphone-array-based sound localization, geometric source separation and post-filtering, (ii) a single-channel template subtraction block and (iii) an ASR block. As the first step of our analysis strategy, we divided the whole-body motion noise problem into three subdomains of arm, leg and head motion noise, according to their intensity levels and spatial location. Subsequently, by following a synthesis-by-analysis approach, we determined the best method for suppressing each type of ego-motion noise. Finally, we proposed to utilize a control module in our ASR framework; this module was designed to make decisions based on instantaneously detected motions, allowing it to switch to the most appropriate method for the current type of noise. This proposed system resulted in improvements of up to 50 points in word correct rates compared with results obtained by single microphone recognition of arm, leg and head motions.

Keywords

Robot audition, noise reduction, whole-body motion noise, automatic speech recognition, template subtraction

[*] To whom correspondence should be addressed. E-mail: gokhan.ince@jp.honda-ri.com

DOI:10.1163/016918611X579448

1. Introduction

Mobile robots are intended to be deployed to environments in which many noise sources are simultaneously present. Therefore, during an interaction with a human, a robot audition system must be able to cope with various kinds of noises, including the robot's own noise (i.e., ego noise). One special type of ego noise, which is observed while the robot is performing an action using its motors, is called ego-motion noise. Owing to the complex characteristics of this particular type of noise, it has so far either been treated with an Stop–Act–Sense loop [1] or circumvented by close-talk microphones [2], because the problem is rather challenging. The complexity of ego-motion noise is further increased when larger numbers of motors are used for a motion, indicating that the noise is even more severe for moving robots with high degrees of freedom. Since mobility is absolutely necessary to improve perceptual capabilities of robots, autonomous robots require robust ego-motion noise suppression abilities at any moment. Due to the increasing popularity of, and the growing demand for, home/service robots, ego-motion noise is likely to become a more significant problem in robotics in the near future.

Although sound source localization and sound source separation problems with background noise or interfering sounds (e.g., human speech, music) have been studied extensively for a long time [3–8], automatic speech recognition (ASR) in the presence of ego-motion noise has received little attention. This type of interference is more difficult to cope with than background noise or static fan noise of the robot, because ego-motion noise is non-stationary and, to some extent, similar to the signals of interest. Therefore, conventional noise reduction methods like spectral subtraction [9, 10] do not work well in practice in reducing ego-motion noise. In addition, the noise sources are present in the near-field of the robot, considerably reducing the performance of conventional far-field noise cancellation methods.

In this study, we propose a method to predict and remove ego-motion noise using templates (discrete audio segments associated with the current motor noise) recorded in advance. Our technique, called parameterized template subtraction, uses templates based on current motor status and the spectral energy vector to represent the ego-motion noise for each time frame at any instance. It incorporates tunable parameters to cope with noise template representations that do not match the instantaneous noise due to deviations in noise spectra. Although this method is effective for removing noise, it suffers from the distorting effects of musical noise [10], similar to all nonlinear single-channel-based noise reduction techniques, and reduces the intelligibility and quality of the audio signal. To also cope with dynamically changing environmental factors, such as background and stationary noise, we apply a nonlinear noise reduction technique for stationary background noise, e.g., minima controlled recursive averaging (MCRA) [9], prior to ego-motion noise reduction. The use of two consecutive nonlinear noise reduction operations (MCRA + template subtraction), however, produces even more musical noise, eventually damaging the acoustic features and reducing the recognition performance of ASR.

To compensate for this effect, we extended the single microphone-based template subtraction method to a hybrid system: (i) a multi-channel noise reduction block consisting of sound source localization (SSL), sound source separation (SSS) and speech enhancement (SE), and (ii) the previously mentioned template subtraction block. While spectral enhancement techniques are the most suitable way to deal with diffuse noise, source separation improves the signal-to-noise ratio (SNR) of the noisy signal by removing directional noise components of ego noise from speech.

In this respect, the first contribution of our work is the integration of the above-mentioned speech processing methods into a single framework to perform ego-motion noise cancellation. Furthermore, we propose an original strategy to solve the whole-body motion noise problem of a robot. Instead of tackling the whole-body motion noise problem holistically, we utilized a synthesis-by-analysis approach. In the first step, the whole-body motion noise problem was partitioned mainly into three ego-motion noise categories: arm, leg and head motion noise, depending on their intensity levels, diffuseness and directivity properties. We assumed that, in a typical interaction scenario, most robots do not use all of their body joints at the same time to perform a certain action, since that would make tasks like coordination of body dynamics, sensor processing and perceptual understanding of the environment much more complicated and difficult. A certain degree of body stationarity (immobility) may improve the reliability of the accomplishment of the task. Even humans, who have very high capabilities of multi-tasking, do the same when focusing on a certain task. While performing task-related motions, humans try to avoid unnecessary motions of the limbs not involved in the task (e.g., keeping the legs fixed while washing dishes). By considering these aspects, we formulated individual solutions to all three types of noise. Multi-channel noise reduction was used to suppress arm and leg noises, because of their highly directional noise characteristics, whereas template subtraction was used to suppress head noise, which is very loud and has complex propagation characteristics. After presenting the results of our analysis, we finalize our architecture by utilizing an ASR module that copes with all kinds of ego-motions. This system was used to select the most appropriate speech features refined by either of the two noise reduction techniques, depending on the type of noise, and to utilize an ASR with the corresponding acoustic model. The proposed system was able to suppress even the noise of motions performed using multiple joints that operate in different categories. We demonstrate that the proposed system improved ASR accuracy.

The rest of the paper is organized as follows. In Section 2, we discuss related work on existing ego-motion noise suppression methods. Section 3 describes the main challenges addressed in this paper and our solutions. In Section 4, we present the details of a template-based ego-motion noise reduction method with spectral enhancement parameters, an extended hybrid system supported by multi-channel noise reduction modules and a switching module for ASR input selection. Section 5 describes our system and its implementation. We present the conducted experiments and demonstrate recognition capabilities of the proposed system in Section 6. We give our conclusions and future work in Section 7.

2. Related Work

Nakadai *et al.* [11] proposed a noise cancellation method using two pairs of microphones. One pair, in the inner part of the shielding body, records only internal motor noise and helps the sound localizer to distinguish between spectral sub-bands that are noisy and not noisy, and to ignore the sub-bands in which noise is dominant. In contrast to our approach, this technique was not designed to remove the noise and obtain refined speech. Its major drawback was that, by filtering out the noise, it also eliminates useful signals. The ego-motion noise problem has also been addressed by predicting and removing ego-motion noise using templates recorded in advance. For example, Nishimura *et al.* [12] estimated the ego-motion noises of distinct gestures of the robot. Using motion commands, the correct noise template matching the corresponding motion was selected from the template database and subtracted. In contrast to their small noise template database of limited and short motions, we targeted the entire repertoire of whole-body motion noise generated by any possible combination of the robot motors. Furthermore, a blockwise template prediction, as in Ref. [12], which uses templates recorded from the onset until the offset time of motor noises, fails completely when the exact onset of the template is not detected properly or the trajectory/duration of the motion changes slightly. Our framewise template prediction method can deal with these problems by representing the motor noise in smaller fragments. Ito *et al.* [13] used an artificial neural network (ANN) to develop a new frame-by-frame-based prediction to cope with unstable walking noise. This ANN also solved the synchronization problem of Nishimura's template-based approach. The trained network was designed to predict the noise spectrum from the angular velocities of the joints of the robot. However, they concentrated on a small robot with limited degrees of freedom. For a huge dataset, ANN will have a slow training speed and online adaptation will be difficult. Therefore, due to its efficiency, we propose using a template database. Approximate search strategies for selecting the appropriate templates make our method more suitable for online learning. In addition, we enhance the accuracy of the templates further by incorporating more information related to the joints, such as angular acceleration. Previous works [12, 13] were based mainly on estimating templates for different motions, but did not focus on the possibility of quality improvement resulting from spectral enhancement optimization factors.

In the field of 'robot audition', noise is suppressed primarily by using sound source separation techniques with microphone arrays [5–8]. Neither a directional noise model, such as that utilized for interfering speakers, nor a diffuse background noise model [6, 7], is entirely appropriate for ego-motion noise. As the motors are located in the near field of the microphones, they produce sounds that have both diffuse and directional characteristics. In a related study, Even *et al.* [14]

proposed to use semi-blind signal separation to obtain both external and internal noise by attaching additional sensors inside the robot. The predictions were used to compute Wiener coefficients. After this suppression step, a delay-and-sum beamformer enhances the refined speech. Although it improves speech recognition accuracy considerably, the additional sensors inside the robot cover pose constraints on implementation. For certain types of sensors, this method may require a body cover made of high-quality or thick material so that external noise is definitely not recorded by these additional sensors (i.e., microphones) or an accurate correspondence model between different sensor signals may be required (i.e., tactile or vib sensors). Besides, semi-blind signal separation methods demonstrate good performance only if the interfering signal is known; however, our estimated noise templates are not sufficiently accurate to be used in a semi-blind signal separation method. In contrast, our method has several advantages, including its ability to be easily implemented on any mobile robot, regardless of the physical constraints about the external shielding. By exploiting only existing microphones, it is also cost-effective and applicable without any hardware modifications.

Several studies have focused on specific conditions for near-field sound sources. For example, Mizumachi *et al.* described a model for sound sources in the near field with spherical wave propagation and line sound sources in contrast to conventional far-field assumptions like plane wave propagation and point-shaped sound sources [15]. Zheng *et al.* proposed a spherically isotropic noise model for near-field objects that achieves stronger reverberation suppression and reduced beampattern variations for broadband signals like our motor noise signals [16]. However, these proposed models are computationally expensive, can only deal with single sound sources, and, more importantly, are designed for stationary sound sources. In a standard task with robot motions, acoustic properties of the noise such as the power and frequencies of the motor noise spectrum as well as the location and number of the active motors, dynamically change over time, thus, reducing its performance considerably.

3. Issues and Approaches

Our goal was to develop a robot audition system that cancels whole-body motion noise of a robot, thus improving the recognition performance. To be able to develop this system, we had to deal with two issues:

(i) *Automatic speech recognition in the presence of ego-motion noise.* Conventional ASR systems assume clean inputs of speech signals. Therefore, a speech signal mixed with ego-motion noise must be refined due to the distorting effects of the noise. If the ego-motion noise is not suppressed, there would be a mismatch between noisy signals and the trained acoustic models resulting in degraded speech recognition performance.

(ii) *Applicability of ego-motion noise reduction to the whole-body motion of a robot*. The robot is expected to use different parts of its body to accomplish certain tasks, including locomotion, object manipulation and object tracking. Based on the complexity of each motion or behavior, the robot may perform several tasks at one time or a task may involve motions of several body parts of the robot, thus involving several joints at one time. We had to confirm whether ego-motion noise could be suppressed regardless of the body part generating the noise. Furthermore, it was necessary to suppress the whole-body motion noise of a robot using a single framework.

We dealt with the above-mentioned issues using the following approaches:

(i) *Ego-motion noise suppression*. We integrated two different methods of ego-motion noise suppression: a template-based ego-motion noise reduction method with spectral enhancement parameters, and a multi-channel noise reduction chain with SSL, SSS and SE stages. The former is suitable for dealing with ego-motion noise problems, because the noise follows a similar pattern each time the respective motion is performed. The templates are good representations of motor noise when the same action was performed during the training phase. The latter method, however, is effective in suppressing directional sound sources. Since the noise originates from the motors that move relative to the positions of the microphones, the noise can be considered directional. Both techniques cannot only be applied individually, but can also be merged into a hybrid system. The output of the noise suppression was used as the input of consequent audio processing stages for various purposes. In this study, we were especially interested in one particular application: ASR.

(ii) *Motion type-based selective ASR module*. Instead of tackling the whole-body motion noise problem holistically, we utilized a synthesis-by-analysis approach. We divided the whole-body motion noise problem mainly into three ego-motion noise categories: arm, leg and head motion noise, depending on the spatial locations of each relative to the microphones and their intensity levels. Figure 1 illustrates three spectrograms of corresponding limb motions, show-

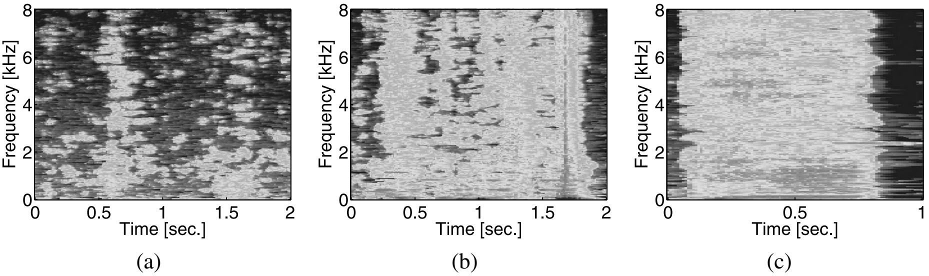


Figure 1. Typical spectrograms of three types of motor noise. (a) Arm motion noise. (b) Leg motion noise. (c) Head motion noise.

ing that they clearly differed in loudness recorded with the microphone array located on top of the robot head (see Section 5 for details about the system). We also observed an uneven energy distribution, with higher noise energy at lower frequencies. These domains are distinctive in being the three main body parts of a robot with end-effectors. To relocate one end-effector, it is usually necessary to relocate a series of joints connected to the end-effector. Therefore, the accumulated noise of the joints can be considered as originating from a certain area. In addition, some tasks can be performed without needing to use all body joints at the same time. Tasks such as the coordination of body dynamics, sensor processing and perceptual understanding of the environment can become complicated and difficult. Since the immobility of distinct body parts unrelated to the task is always desired to improve the reliability of accomplishing the task, we provide individual solutions for all three types of noise. We also assessed the performance of ASR for all three ego-motion noises and their combinations. After presenting the results of our analyses, we propose a final architecture that can deal with all types of ego-motions and their combinations (whole-body motion), using a motion type-based selective ASR module with a switching mechanism for ASR inputs.

4. Proposed Methods

In this section, we describe the theoretical background of the basic building blocks of the proposed system architecture. Section 4.1 describes the details of the template subtraction method. Section 4.2 illustrates how single-channel template subtraction can be extended into a hybrid framework to suppress ego-motion noise by incorporating existing multi-channel noise reduction techniques.

4.1. Single-Channel Template Subtraction

The underlying motivation of using templates for noise reduction resides in the fact that the duration of motor noise signals is similar for the same motions performed repeatedly and the envelope of the signal does not deviate much from the mean envelope. However, conventional blockwise template subtraction [12], which uses templates recorded from the start to the end of motor noises, has several shortcomings, including the need to start subtraction only after the detection of the exact starting moment of the template — a task very hard to achieve. Another drawback of this method is its requirement of a large collection of signal representations of every possible motion trajectory, consisting of motor noise statistics, such as averages and standard deviations. Moreover, this method requires a huge amount of data for each possible motion. Due to the impossibility of collecting and producing templates for each joint for each combination of origin, target, position, velocity and acceleration parameters, this approach is simply not feasible in realistic scenarios on humanoid robots with free motion selection.

To overcome these deficits, a new technique was implemented [17], which parameterizes discrete audio segments using motor status and results in a spectral energy vector representing the ego-noise at that instant. To implement this so-called parameterized template subtraction, we needed a robot with joint angle sensors (encoders) that measure the angular positions of each of its joints separately. In addition, this method can improve the quality of speech using spectral enhancement parameters (see Section 4.1.3).

Before explaining the details of parameterized template subtraction, we define $S(\omega, k)$ and $D(\omega, k)$ as the short-time cross-correlation spectra of useful signal and distortion (motor noise only), respectively, where ω represents the discrete frequency and k represents the time-frame. Thus, the spectrum of the observed signal $X(\omega, k)$ can be described as:

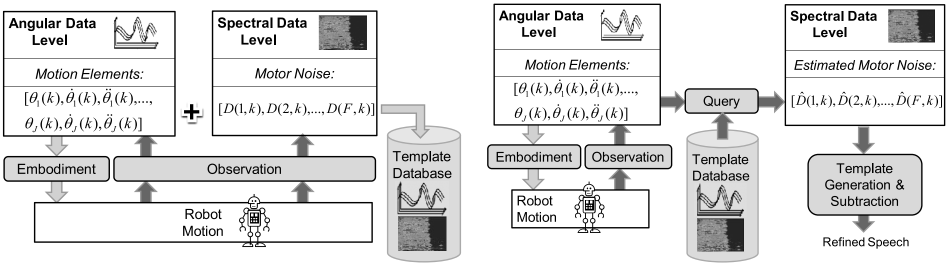
$$X(\omega, k) = S(\omega, k) + D(\omega, k). \tag{1}$$

4.1.1. Template Database Generation

We utilize joint status information provided by the sensors on the motors under the following assumptions:

- The noise of a motor is dependent on the position, velocity and acceleration of that motor.
- Similar combinations of joint status will result in similar motor noise spectral vectors at any instant of time.
- The superposition of single joint motor noises at an arbitrary time point is equal to the whole-body noise at the corresponding time point.

Figure 2a shows the proposed template database generation scheme. During the motion of the robot, the actual position (θ) of each motor is gathered regularly. Using the difference between consecutive sensor outputs, velocity ($\dot{\theta}$) and acceleration ($\ddot{\theta}$) are calculated. If J joints are active, $3J$ features will be generated. Each feature is normalized to [-1 1], so that all features make the same contribution to the prediction. The resulting feature vector is in the form, $[\theta_1(k), \dot{\theta}_1(k), \ddot{\theta}_1(k), \ldots, \theta_J(k), \dot{\theta}_J(k), \ddot{\theta}_J(k)]$. At the same time, motor noise is recorded and background noise is removed from the recordings. The spectrum of


Angular Data Level
Motion Elements:
$[\theta_1(k),\dot{\theta}_1(k),\ddot{\theta}_1(k),...,$
$\theta_J(k),\dot{\theta}_J(k),\ddot{\theta}_J(k)]$
+
Spectral Data Level
Motor Noise:
$[D(1,k),D(2,k),...,D(F,k)]$
Embodiment
Observation
Robot Motion
Template Database
Angular Data Level
Motion Elements:
$[\theta_1(k),\dot{\theta}_1(k),\ddot{\theta}_1(k),...,$
$\theta_J(k),\dot{\theta}_J(k),\ddot{\theta}_J(k)]$
Query
Spectral Data Level
Estimated Motor Noise:
$[\hat{D}(1,k),\hat{D}(2,k),...,\hat{D}(F,k)]$
Embodiment
Observation
Robot Motion
Template Database
Template Generation & Subtraction
Refined Speech

(a) (b)

Figure 2. Proposed template database generation and template prediction method for ego noise suppression. (a) Flowchart of template database generation. (b) Flowchart of template prediction.

the motor noise ($[D(1,k), D(2,k), \ldots, D(F,k)]$, where F represents the number of frequency bins) is calculated by the sound processing branch running in parallel. Both feature vectors and spectra are continuously labeled with time tags so that the corresponding templates are generated when their time tags match. Finally, a large database of noise templates consisting of one-frame noise templates for many joint configurations is created.

4.1.2. Motor Noise Prediction

The prediction phase starts with a search in the database for the motor noise template that best matches the motor noise at that point in time (Fig. 2b). Finding the correct template involves a search among all the templates in the database for the most similar joint configuration. We implemented a nearest-neighbor ($1-\text{NN}$) search to accomplish this task. The spectral vector associated with the point in the database with the shortest distance to the query point was selected as the template. The prediction process is repeated for each time frame. In that sense, the conventional 'blockwise template' for a single arbitrary motion can be regarded as the concatenation of smaller templates that are predicted according to the abovementioned approach on a frame-by-frame basis.

4.1.3. Template Generation and Subtraction

The spectrum of the useful signal in (2) can be obtained using the inverse operation of (1):

$$S_{\mathrm{r}}(\omega,k) = X(\omega,k) - \hat{D}(\omega,k), \tag{2}$$

where $\hat{D}(\omega,k)$ denotes the estimated noise template and $S_{\mathrm{r}}(\omega,k)$ represents the signal that includes both the useful sound and the residual motor noise. The presence of this residual noise is due to a deviation between the original motor noise $D(\omega,k)$ and the predicted motor noise $\hat{D}(\omega,k)$. To compensate for this error, we utilized a spectral subtraction approach that encompasses both an overestimation factor, α, and a spectral floor, β. The parameter α, also called an aggressiveness factor, allows a compromise between perceptual signal distortion and noise reduction level. In contrast, β is required to deal with the problem called musical noise, which is caused by nonlinear mapping of the negative or small-valued spectral estimates. This produces a metallic noise, which sounds like the sum of tone generators with random fundamental frequencies that are constantly turned on and off [9]. The parameter β reduces the effects of the sharp valleys and peaks in the spectrum caused by smaller attenuations of frequencies compared with the relatively larger attenuations of their neighboring frequencies due to random fluctuations in magnitude estimations. Finally, parameterized template subtraction can be introduced using the formula:

$$\hat{H}_{\mathrm{SS}}(\omega,k) = \max\left(1 - \alpha\frac{|\hat{D}(\omega,k)|}{|X(\omega,k)|}, \beta\right), \tag{3}$$

where $\hat{H}_{\mathrm{SS}}(\omega, k)$ represents the gain coefficient and $|\cdot|$ represents the magnitude spectra of the signals. A weighting operation of the signal $X(\omega, k)$ with this coefficient finalizes the template subtraction as:

$$\hat{S}(\omega, k) = \hat{H}_{\mathrm{SS}}(\omega, k) \cdot |X(\omega, k)| \cdot \mathrm{e}^{j \arg(X(\omega,k))}. \tag{4}$$

In contrast to previous methods in Refs [12, 13], our prediction, generation and subtraction methods do not require any starting or ending signals. Thus, there are no abrupt blockwise templates applied discontinuously to the noisy signals. Our methods process data continuously, even when the robot does not move. Therefore, our template database does not only consist of recordings of motor noise, but also of recordings of the joints in resting positions. Each training session consisted of uninterrupted sound recordings of a single continuous motion sequence consisting of hundreds of consequent motions with short (less than 1 s) pauses between each motion.

4.2. Hybrid Framework with a Switching Module for ASR Input Selection

Section 4.2.1 briefly discusses each module (SSL, SSS and SE) of the multi-channel noise reduction block that we incorporated into our system. Further details are provided by the given references published in various robotics journals and conferences. In addition, the modules described in Section 4.2.1 can be replaced by other multi-channel solutions capable of separating directional sound sources. In Section 4.2.2, we explain the key module of our proposed architecture that results in the hybrid system.

4.2.1. Multi-Channel Noise Reduction System [18]

To estimate the directions of arrival (DoA) of each sound source, we used a popular adaptive beamforming algorithm called MUltiple SIgnal Classification (MUSIC) [19]. This algorithm detects each DoA by performing eigenvalue decomposition on the correlation matrix of the noisy signal, by separating subspaces of undesired interfering sources and sound sources of interest, and finally by identifying the peaks occurring in the spatial spectrum. A consequent source tracker system performs temporal integration in a given time window.

Geometric source separation (GSS) [20], later extended to an adaptive algorithm that can process the input data incrementally [7], makes explicit use of source locations to separate different sound sources. To properly estimate the separation matrix, GSS introduces cost functions that must be minimized in an iterative way [7]. Moreover, we used adaptive step-size control, resulting in fast convergence of the separation matrix [21]. Our GSS implementation also exploited a method called optima controlled recursive averaging, which controls the window size adaptively, causing smoother convergence and, thus, better separation results [5]. Specifically, GSS has three distinct advantages for the ego-noise cancellation problem.

(i) The introduction of the concept of geometric constraints, which involves calculations of current transfer functions based on the known locations of the microphones and the positions of the sound sources obtained from SSL. This relaxes the limitations of Blind Source Separation (BSS), such as permutation and scaling problems, and can, therefore, run in real-time.

(ii) Sound separation of moving sources is possible. This is especially important since the part of the robot on which the microphones are mounted (e.g., the head) can also move. Relative to a moving microphone array, even stationary sound sources are regarded as moving objects.

(iii) Generally, an embodied robot has loud ego noises, such as the stationary operational noise of hardware and fan noise, which are located close to each other. If the positions of these high noise emission sources are known, their directions can be specified, because our GSS module has a function that suppresses stationary ego noise as a fixed noise source.

The separation process is followed by a multi-channel post-filtering operation, enabling the sounds to be enhanced further. This module was based on the optimal estimator proposed by Ephraim and Malah [22]. Since their method takes temporal and spectral continuities into consideration, it generates less distortion than conventional spectral subtraction-based noise reduction methods. By further extending this idea, we were able to apply a multi-channel post-filter [7], which can cope with non-stationary interferences as well as stationary noise. This module treats the transient components in the spectrum as if they are caused by leaking energies that may occasionally arise due to poor separation performance. For this purpose, noise variances of both stationary noise and source leakage are predicted, with the former computed using the MCRA [10] method and the latter estimated using the algorithm proposed in Ref. [7]. The noise suppression rule also involves speech presence probability calculations [23] and is based on the minimum mean-square error estimation of the spectral amplitude [22].

4.2.2. *Switching Module for ASR Input Selection*

After initially analyzing the performances of multi-channel and single-channel noise reduction methods (see Section 6), in the synthesis stage, we suggested processing the speech feature outputs of both pathways in a single ASR module and to use them interchangeably in a motion-dependent fashion, as in Fig. 3. This switching module is triggered by the output of the motion detector. As this system gathers information about all joints at every moment of time, the switching module is able to discriminate among joints that are and are not actively involved in the motion by checking their velocities, and can, therefore, determine the motion being performed at that moment. The module then switches between the outputs of single-channel and multi-channel noise reduction-based speech features. As the multi-channel approach works better than the single-channel approach for the leg and arm noises (Section 6.2), the switch feeds the acoustic features of this branch to the ASR module whenever a leg and/or an arm motion is detected. It also utilizes the acoustic model trained for the multi-channel approach. For a head motion, however, features

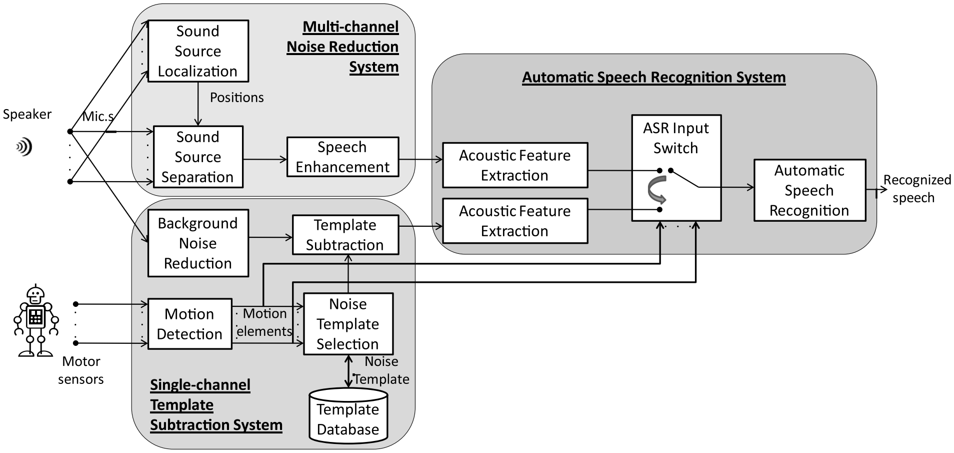

Multi-channel Noise Reduction System
Sound Source Localization
Positions
Sound Source Separation
Speech Enhancement
Speaker
Mic.s
Background Noise Reduction
Template Subtraction
Motion Detection
Motion elements
Noise Template Selection
Noise Template
Template Database
Motor sensors
Single-channel Template Subtraction System
Automatic Speech Recognition System
Acoustic Feature Extraction
Acoustic Feature Extraction
ASR Input Switch
Automatic Speech Recognition
Recognized speech

Figure 3. Proposed noise cancellation system.

generated after template subtraction are more suitable (Section 6.3). We observed similar recognition performance during simultaneous head and arm motion noises (Section 6.3), indicating that head motion contributes to whole-body motion noise more than any other motion from other domains. In summary, as long as head motion is present, we can suppress whole-body motion noise using single-channel noise reduction methods; otherwise the multi-channel noise reduction method is used. We implemented the following rule-based routing in the switch:

$$\text{Decision}(k) = \begin{cases} \text{Acoustic features of single-channel} & \\ \quad \text{template subtraction,} & \text{if any } |\dot{\theta}_{\text{HeadJoint}}(k)| > \varepsilon \\ \text{Acoustic features of multi-channel} & \\ \quad \text{noise reduction,} & \text{otherwise,} \end{cases} \tag{5}$$

where $|\dot{\theta}_{\text{HeadJoint}}(k)|$ denotes the absolute velocity of the pan or tilt motion of the head and ϵ is a speed threshold. We proposed to use ε, instead of zero, to prevent the activation of the switch during the tail motion of the head; it is used as a countermeasure to situations during which motion has stopped, but the joint sensors continue to send very small differences in position.

5. System Architecture and Implementation

The overall architecture of the proposed noise reduction system consisting of three blocks is shown in Fig. 3. The first block (Multi-Channel Noise Reduction System) starts with a module performing SSL and extracting the location of the most dominant sources in the environment. The estimated locations of the sources are processed by the linear separation algorithm, SSS, followed by a SE step. This module attenuates stationary noise (e.g., background noise) and non-stationary noise that arises because of the leakage energy between the output channels of the previous separation stage for each individual sound source (speaker and directional fan noise). The second block performs single-channel template subtraction [17]. Together, both blocks are responsible for producing spectrograms for the extraction of audio features in the last block, which performs ASR. The whole system (including joint status acquisition, sound recording and sound processing stages) runs synchronously based on a single clock. Data flow is realized mainly by means of fixed-length audio frames.

This system was specifically designed for single-speaker speech recognition tasks. This means that we have assumed that there is no directional sound source other than the speaker, whose speech would be recognized. Therefore, external noise is considered only as background noise, which is a diffuse noise by its nature. Both branches of the hybrid system can deal with diffuse noise by utilizing either background noise reduction (single-channel template subtraction) or MCRA inside SE (multi-channel noise reduction). To tackle the motor noise, on the other hand, we use either template subtraction (single-channel noise reduction for head motion) or SSS (multi-channel noise reduction for arm and leg motion). Thus, both branches can deal with internal and external noise in their own ways.

6. Evaluation

To evaluate the performance of the proposed techniques, we used a humanoid robot developed by Honda. This robot is equipped with an eight-channel microphone array on top of its head. We used two motors for head motion, five motors to move each leg and four motors to move each arm, resulting in a total of 20 d.o.f. Relative to the microphone array configuration, the neck motors are the closest sound sources, making them the most problematic, because the intensity of a sound wave depends on its distance from its source:

$$\text{Sound Intensity} = \text{Sound Power}/(4\pi R^2), \tag{6}$$

where R denotes the distance. In addition, since all limbs of the robot operate in different, non-overlapping coverage areas, which also help in differentiating noise types by their spatial locations, we decided to handle the noise problem in different domains, each covering a set of joints required for a certain type of interaction with the robot's environment. We, therefore, recorded motions performed by a given set of limbs, which could be classified into three distinct categories, arm motion, leg motion and head motion, in order of increasing noise intensity.

Based on these conditions, we present the experimental settings in Section 6.1. Afterwards, we assess the performance of the noise reduction methods for arm and leg motion noise (Section 6.2) and head motion noise (Section 6.3).

6.1. Experimental Settings

We recorded (i) random, whole-arm pointing motion within the reaching space of the robot body as arm motion, (ii) stamping behavior and short distance walking as leg motion, and (iii) random head rotation (elevation = [$-30°$ 30°], azimuth = [$-90°$ 90°]) as head motion. The average noise energies of leg and head motion were 5.1 and 8.4 dB higher, respectively, compared with those of arm motions. For the second part of the experiments involving template subtraction, we recorded two additional sets of random motions (performed by the head only and by the head and arm together) and stored a training database of 30 min and a test database 10 min long (due to software constraints the joint positions of the legs cannot be acquired; therefore, we could not apply the template subtraction method to leg motion noise). Sensors determine the angle of the joints every 5 ms, with each audio frame being 10 ms in length. We used constant values for $\alpha = 1$ and $\beta = 0.5$ as template subtraction parameters, because we previously observed that, compared with $\beta = 0$, increased β improves ASR accuracy considerably (for detailed evaluations regarding α and β, and their effects on ASR accuracy, signal quality and noise suppression rates, see Ref. [17]).

As the noise recordings are longer than the utterances used in isolated word recognition, we selected those segments in which all joints contributed to noise. To generate precise amounts of noise and speech energy for various SNR conditions before mixing them, we amplified clean speech based on its segmental SNR, *segSNR*. The *segSNR* estimates the SNR level within each segment and averages it over the entire signal, providing a better representation of energy distribution for speech and noise within the relevant time interval under consideration:

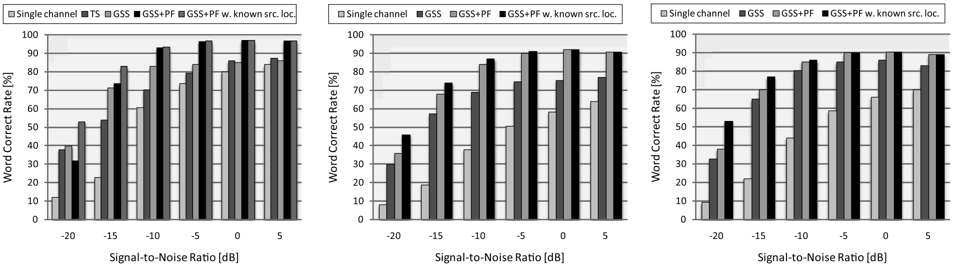
$$segSNR = \frac{1}{J}\sum_{j=1}^{J} 10\log_{10}\left(\frac{\sum_n s_j^2(n)}{\sum_n d_j^2(n)}\right), \tag{7}$$

where J is the number of segments with speech activity, and $s(n)$ and $d(n)$ are the nth discrete speech and noise samples, respectively. The noise signal, consisting of ego noise (including ego-motion noise) and environmental background noise, was mixed with clean speech utterances used in a typical human–robot interactive dialog. This Japanese word dataset includes 236 words for four female and four male speakers. Acoustic models were trained with the Japanese Newspaper Article Sentences corpus, 60 h of speech data spoken by 306 male and female speakers, making speech recognition a word-open test. The results for template subtraction (TS) were evaluated using an acoustic model trained with MCRA-applied speech data. In contrast, we used a matched acoustic model for multi-channel noise reduction (GSS+PF) methods. Both of these models were trained with data processed at motor noise conditions of SNR levels ranging from -10 to 5 dB. We used 13 static mel-scale log spectrum (MSLS) [24] features, 13 delta MSLS features and one delta power feature. Speech recognition results are given as average word correct rates (WCR) of instances from the noisy test set. The position of the speaker was kept fixed at 0° throughout the experiments. The recording environment consisted of a 4.0 m × 7.0 m × 3.0 m room with a reverberation time (RT_{20}) of 0.2 s. The implementation was run on HARK — an open-sourced software for robot audition [25].

6.2. Speech Recognition with Arm and Leg Motion Noise

In this experimental setting, the microphone array and the head were kept stationary, allowing us to fix the direction of the ego noise (fan noise) originating from the backpack of the robot at −180°. Providing a fixed ego-noise direction did not pose any hard constraints on robot audition scenarios or applications, because the robot was already equipped with sensors to transmit the positions of the joints. Depending on the posture of the body, we were able to determine exactly the source of the ego noise and transmit the direction automatically to our source separation algorithm as input. We present the results for GSS and GSS+PF where the position of the speaker was detected using our implementation. As an additional test, we also determined 'GSS+PF with known source location' results — a condition where we assumed that the location of the sound source was estimated precisely.

Figure 4 shows speech recognition accuracies for arm, leg, and arm plus leg motions at the same time. Single-channel results without processing were used as baseline. Template subtraction resulted in good ASR accuracy, but its performance was inferior to that of GSS+PF (TS evaluation of leg motion was not possible). Under all three conditions, the multi-channel noise reduction system resulted in an up to 40 points improvement compared with single-microphone-based recognition. In general, these results indicate that the directional effects of arm and leg motions noise can be treated with GSS, and that residual noise (as diffuse components) can be partially handled by PF. As the arms operate mostly on the right- and left-hand sides of our humanoid robot, their noises can be separated well due to the spatial (angular for GSS) distance between the arms and the target speaker standing in front of the robot. In addition, the leg noise came from below the waist of the robot, making its distance from the microphone array large enough for separating it from the speaker. As long as the direction of the ego-motion noise is not the same as that of the target speaker, this method works well to suppress all ego noise, both ego-motion noise and fan noise. Furthermore, the recognition result curves in Fig. 4c show very similar patterns to the curves in Fig. 4a and b. This very promising result



(a) (b) (c)

Figure 4. Recognition performance during arm and leg motions. (a) Arm motion noise. (b) Leg motion noise. (c) Arm plus leg motion noise.

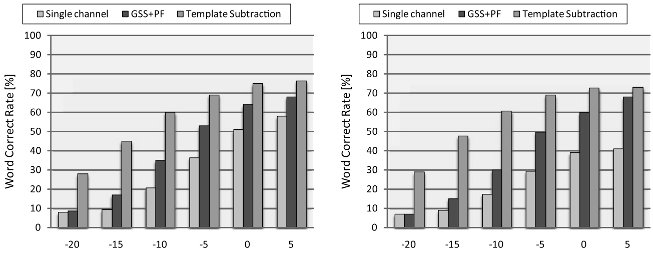
indicates that GSS+PF is effective even when it is used against the combination of arm and leg motion noise.

It is also noteworthy to mention that SSL fails in low SNRs due to its fixed threshold operation. SSL estimates additional non-existing ‘ghost’ sources, decreasing the performance of GSS and PF. In contrast, GSS+PF with known source location demonstrates the upper performance limit of our proposed method.

6.3. Speech Recognition with Head Motion Noise

One consequence of head motion is the relative motion of sound sources with respect to the microphones. Whenever the head moves, the microphone array also moves. Since we tested only isolated word recognition, we hypothesized that the effects of the moving sound sources on separation and speech enhancement performance were rather small, but in fact not negligible. Nevertheless, to inspect the capabilities of our proposed noise reduction system based on SSS, we did not provide the ego-noise direction of the robot in advance; rather, the SSL system predicted it automatically.

Figure 5a illustrates the ASR accuracy for head motion noise. The multi-channel approach provided poorer performance than the single-channel template subtraction technique, because short-range reverberation effects and multi-path propagation inside/outside the head are properties of head motion noise that are very hard to overcome with the current GSS+PF algorithm limits and settings. The neck motors are located inside the head cover, where the microphones are also installed. As head motor noise propagates inside the head in a highly reverberant way in close proximity to the microphones, the directional noise assumption is violated. Strong noise sources in the very near field of the microphone array have highly complicated propagation patterns. As a consequence, it worsens the separation quality. Thus, the noise model used in the post-filtering is not applicable under these conditions. TS resulted in better improvement, because it does not model the noise depending on its directivity–diffuseness nature, but rather instantaneously predicts the current noise template from a database, depending on the position and velocity of the joints. In addition to being prone to modeling errors, it also suffers from musical noise com-

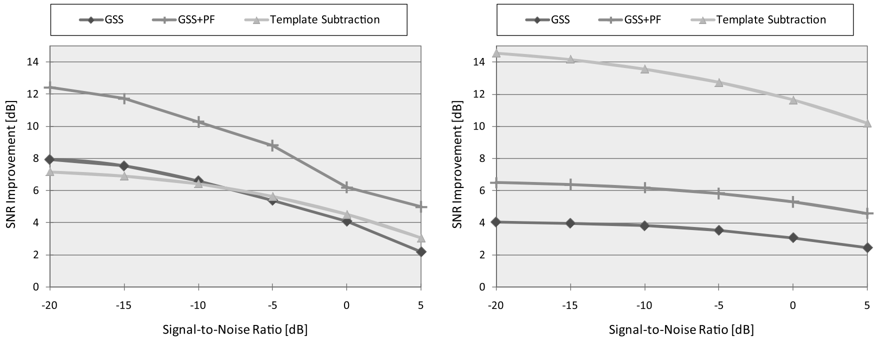


(a) (b)

Figure 5. Recognition performance during head motion. (a) Head motion noise. (b) Head plus arm motion noise.

Table 1.
Recognition accuracy (%) for different ego-motions achieved by single-channel template subtraction and multi-channel noise reduction (SNR = −5 dB)

	Arm	Leg	Arm + leg	Head	Head + arm
Single channel	73	50	58	37	30
Template subtraction	80	—	—	**69**	**69**
Multi-channel noise reduction	**96**	**90**	**90**	53	50



(a) (b)

Figure 6. SNR improvements of the compared methods. (a) During arm motion noise. (b) During head motion noise.

ponents caused by subtraction in the spectral domain. This distorts the spectrum and degrades features, making the WCR improvement rather limited, but still better than GSS+PF.

Table 1 shows that the proposed hybrid system highly improves the elimination of all motor noise types and their combinations within the limits of our implementation conditions. Although Table 1 presents results for only one moderate SNR level of −5 dB, the trend of improvement was true for all SNRs used in these experiments. Our switching module always selects the noise suppression method that yields the best performance for the undertaken action.

Finally, to assess the suppression capabilities of both methods, we also assessed SNR improvement by calculating the differences in SNR before and after the application of noise cancellation methods. Mean SNR improvements are shown in Fig. 6. For head motion noise the template subtraction method resulted in higher suppression performance than GSS and GSS+PF for head motion noise, whereas GSS+PF reduces the largest portion of arm motion noise. Although the trends of SNR improvements were consistent with the WCR curves, high suppression rates do not necessarily indicate higher recognition accuracy for TS as long as noise templates are not correctly estimated.

7. Conclusion

We have described methods for eliminating whole-body motion noise from speech signals. Since ego-motion noises arising from the motors of a robot are created in the near field of the microphone array, these noises have both diffuse and directional characteristics. We used a synthesis-by-analysis approach to suppress this noise. We divided whole-body motion noise into three domains, depending on their spatial location and intensity levels: arm, leg and head motion noise. The system we proposed extracts information about the motion type performed at that moment and decides on the best choice of processing method for speech recognition by a selective ASR module. We adopted two methods — multi-channel noise reduction and single-channel template subtraction — which are switched depending on the detection of the head motion. If no head motion is detected, the first method is selected because it is effective for arm, leg, and arm plus leg motion noises. This method utilizes SSL incorporating the MUSIC algorithm and SSS using the GSS algorithm, finalized by a SE stage that suppresses both background noise and interference/leakage noise. Source separation is particularly effective against noises from the arms and legs, because the limbs are located away from the microphone array and are separated from a speaker standing directly in front of the robot. On the other hand, if head motion is detected, the second method is selected. It is more appropriate for canceling head motion noise (or the combination of the head motion noise with arm and/or leg motion noise), because template subtraction makes no assumptions about the nature of the noise and uses previously recorded noises. We validated the applicability of our approach by evaluating its performance on three different motor noise types and their combinations. Our method demonstrated good performance in suppressing arm, leg and head motion noise, and their combinations, as shown by ASR accuracy.

The optimal system structure for ego noise suppression depends solely on the characteristics of the ego noise, as long as the ego noise is not picked up by additional sensors and must be estimated. We found that single-channel noise reduction was far superior to multi-channel noise reduction in suppressing motor noise recorded by closely located microphones. Owing to the complex characteristics of motor noise propagation inside the robot cover, where the microphones are mounted, blind source separation and speech enhancement perform very poorly in these conditions (i.e., head motion noise). In contrast, when the motors are located further from the speaker and microphones, both in distance and separation angle, the multi-channel noise reduction was more effective (i.e., arm plus leg motion noise). Therefore, the proposed parallel system architecture can be considered optimal for any robotic system containing only robot-embedded microphones and with the switch trigger design based on the locations of the microphones and motors.

We also investigated alternative combinations, including a cascaded version of SSS+PF+TS, instead of our hybrid architecture. In practice, however, it was not possible to create a template database for ego noise after the SSS+PF stages. One reason is that the recording must be performed only when there is no external directional sound source, but performing an SSS without any sound sources is not possible. In addition, the template database must contain ego noise artifacts after SSS+PF for all directions in which a candidate target sound source may be present. It is almost impossible to create such a huge template database. We also evaluated this combination in reverse order, PF+TS+SSS, but it yielded far worse results, because the spectral subtraction prior to SSS damaged the spectra of the microphone signals, resulting in poorer performance of the SSS.

By changing the posture of the robot, so that its body is aimed directly at the target speaker, the robot can avoid the interference due to the ego-motion noise of the arms with the target speaker's utterances by maximizing their spatial distance. Our system remains open for improvements. To apply template subtraction to leg noise, we plan to make changes that will allow us to gather angular information from the legs. One weakness of the current architecture is the fixed threshold operation used in the SSL procedure, which determines if a source is present at that location. As motor noise increases, the system becomes more susceptible to the threshold value. Since no optimal threshold is effective for every kind of motor noise, we plan to make it adaptive. Our multi-channel system in its current form can also deal with four speakers. The next step is to design a system in real-time and in a real situation involving speech recognition of several speakers simultaneously while the robot is performing some motion.

References

1. T. Rodemann, M. Heckmann, B. Schölling, F. Joublin and C. Goerick, Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping, in: *Proc. IEEE/RSJ Int. Conf. on Robots and Intelligent Systems*, Beijing, pp. 860–865 (2006).
2. S. E. Levinson, W. Zhu, D. Li, K. Squire, R. S. Lin, M. Kleffner, M. McClain and J. Lee, Automatic language acquisition by an autonomous robot, in: *Proc. Int. Joint Conf. on Neural Networks*, Portland, OR, pp. 2716–2721 (2003).
3. 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, in: *Proc. IEEE/RSJ Int. Conf. on Robots and Intelligent Systems*, Sendai, pp. 2404–2410 (2004).
4. H. Saruwatari, Y. Mori, T. Takatani, S. Ukai, K. Shikano, T. Hiekata and T. Morita, Two-stage blind source separation based on ICA and binary masking for real-time robot audition system, in: *Proc. IEEE/RSJ Int. Conf. on Robots and Intelligent Systems*, Edmonton, pp. 209–214 (2005).
5. K. Nakadai, H. Nakajima, Y. Hasegawa and H. Tsujino, Sound source separation of moving speakers for robot audition, in: *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Taipei, pp. 3685–3688 (2009).
6. S. Yamamoto, K. Nakadai, M. Nakano, H. Tsujino, J. M. Valin, K. Komatani, T. Ogata and H. G. Okuno, Real-time robot audition system that recognizes simultaneous speech in the real world, in: *Proc. IEEE/RSJ Int. Conf. on Robots and Intelligent Systems*, Beijing, pp. 5333–5338 (2006).
7. J.-M. Valin, S. Yamamoto, J. Rouat, F. Michaud, K. Nakadai and H. G. Okuno, Robust recognition of simultaneous speech by a mobile robot, *IEEE Trans. Robotics* **23**, 742–752 (2007).

8. T. Takahashi, S. Yamamoto, K. Nakadai, K. Komatani, T. Ogata and H. G. Okuno, Soft missing-feature mask generation for simultaneous speech recognition system in robots, in: *Proc. Int. Conf. on Spoken Language Processing (Interspeech)*, Brisbane, pp. 992–997 (2008).
9. S. Boll, Suppression of acoustic noise in speech using spectral subtraction, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-27**, 113–120 (1979).
10. I. Cohen, Noise estimation by minima controlled recursive averaging for robust speech enhancement, *IEEE Signal Process. Lett.* **9**, 12–15 (2002).
11. K. Nakadai, H. G. Okuno and H. Kitano, Humanoid active audition system improved by the cover acoustics, *Lecture Notes Artif. Intell.* **1886**, 544–554 (2000).
12. Y. Nishimura, M. Nakano, K. Nakadai, H. Tsujino and M. Ishizuka, Speech recognition for a robot under its motor noises by selective application of missing feature theory and MLLR, in: *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, Pittsburgh, PA, pp. 53–58 (2006).
13. A. Ito, T. Kanayama, M. Suzuki and S. Makino, Internal noise suppression for speech recognition by small robots, in: *Proc. Int. Conf. on Spoken Language Processing (Interspeech)*, Lisbon, pp. 2685–2688 (2005).
14. J. Even, H. Sawada, H. Saruwatari, K. Shikano and T. Takatani, Semi-blind suppression of internal noise for hands-free robot spoken dialog system, in: *Proc. IEEE/RSJ Int. Conf. on Robots and Intelligent Systems*, St Louis, MO, pp. 659–663 (2009).
15. M. Mizumachi and S. Nakamura, Passive subtractive beamformer for near-field sound sources, in: *Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop*, Barcelona, pp. 74–78 (2004).
16. Y. R. Zheng, R. A. Goubran and M. El-Tanany, A nested sensor array focusing on near field targets, *Proc. IEEE Sensors* **2**, 843–848 (2003).
17. G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino and J. Imura, Ego noise suppression of a robot using template subtraction, in: *Proc. IEEE/RSJ Int. Conf. on Robots and Intelligent Systems*, St Louis, MO, pp. 199–204 (2009).
18. G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino and J. Imura, A hybrid framework for ego noise cancellation of a robot, in: *Proc. IEEE/RSJ Int. Conf. on Robotics and Automation*, Anchorage, AK, pp. 3623–3628 (2010).
19. R. Schmidt, Multiple emitter location and signal parameter estimation, *IEEE Trans. Antennas Propagat.* **34**, 276–280 (1986).
20. L. C. Parra and C. V. Alvino, Geometric source separation: merging convolutive source separation with geometric beamforming, *IEEE Trans. Speech Audio Process.* **10**, 352–362 (2002).
21. H. Nakajima, K. Nakadai, Y. Hasegawa and H. Tsujino, Adaptive step-size parameter control for real-world blind source separation, in: *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Las Vegas, NV, pp. 149–152 (2008).
22. Y. Ephraim and D. Malah, Speech enhancement using minimum mean-square error short-time spectral amplitude estimator, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-32**, 1109–1121 (1984).
23. I. Cohen and B. Berdugo, Microphone array post-filtering for non-stationary noise suppression, in: *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Orlando, FL, pp. 901–904 (2002).
24. Y. Nishimura, T. Shinozaki, K. Iwano and S. Furui, Noise-robust speech recognition using multi-band spectral features, in: *Proc. 148th Acoustical Society of America Meet.*, San Diego, CA, 1aSC7 (2004).

25. K. Nakadai, H. Okuno, H. Nakajima, Y. Hasegawa and H. Tsujino, An open source software system for robot audition HARK and its evaluation, in: *Proc. IEEE–RAS Int. Conf. on Humanoid Robots*, Daejeon, pp. 561–566 (2008).

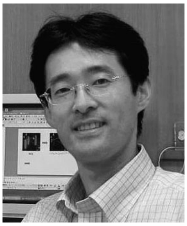
About the Authors

Gökhan Ince received the BS degree in Electrical Engineering, from Istanbul Technical University, Turkey, in 2004, and the MS degree in Information Engineering from Darmstadt University of Technology, Germany, in 2007. He is currently pursuing a PhD degree in the Department of Mechanical and Environmental Informatics, Tokyo Institute of Technology, Japan. From 2006 to 2008, he was a Researcher with Honda Research Institute Europe, Offenbach, Germany. Since 2008, he has also been with Honda Research Institute Japan, Co., Ltd, Saitama, Japan. His current research interests include human–robot interaction, audio processing and auditory scene analysis. He is a Member of the IEEE, RAS, ISAI and ISCA.

Kazuhiro Nakadai received the BE degree in Electrical Engineering, in 1993, the ME degree in Information Engineering, in 1995, and the PhD degree in Electrical Engineering, in 2003, all from Tokyo University. He worked with Nippon Telegraph and Telephone and NTT Comware Corp. as a System Engineer, from 1995 to 1999. He was a Researcher at Kitano Symbiotic Systems Project, ERATO, Japan Science and Technology Agency, from 1999 to 2003. He is currently a Principal Researcher for Honda Research Institute Japan, Co., Ltd. Since 2006, he has also been Visiting Associate Professor at Tokyo Institute of Technology. His research interests include AI, robotics, signal processing, computational auditory scene analysis, multi-modal integration and robot audition. He is a Member of the RSJ, JSAI, ASJ and IEEE.

Tobias Rodemann studied Physics and Neuro-informatics at the Ruhr Universität Bochum, Germany, and received his Dipl.-Phys. degree from the Universität Bochum, in 1998, and a PhD degree from the Technische Universität Bielefeld, Germany, in 2003. Since 1998, he has been working at the Honda Research Institute Europe, Offenbach, Germany. Previous research fields were evolutionary algorithms, biologically inspired vision systems, in formation processing with spiking neurons and learning of sensory-motor maps. Since 2003, he has been working as a Senior Scientist on sound localization, auditory scene analysis and audio–visual interaction.

Hiroshi Tsujino is a Chief Researcher at the Honda Research Institute Japan, where he directs the associative interacting intelligence involving brain-like computing, brain–machine interface and human–robot interaction. He received his MS degree in Computer Science from the Tokyo Institute of Technology, in 1986. In 1987, he joined the Honda Research and Development Co., Ltd, and was engaged in researching intelligent assistance systems for cars, image understanding systems and brain-inspired reasoning systems. In 2003, he joined the Honda Research Institute Japan when it was established. His research focuses on creating intelligent machines that have interacting intelligence with humans in real-world situations. He is a Member of the IEEE, INNS, SFN, JSAI, JSSST and RSJ.



Jun-ichi Imura received the MS degree in Applied Systems Science, and the PhD degree in Mechanical Engineering from Kyoto University, Japan, in 1990 and 1995, respectively. He served as a Research Associate at the Department of Mechanical Engineering, Kyoto University, from 1992 to 1996, and as an Associate Professor at the Division of Machine Design Engineering, Faculty of Engineering, Hiroshima University, from 1996 to 2001. From May 1998 to April 1999, he was a Visiting Researcher at the Faculty of Mathematical Sciences, University of Twente, The Netherlands. Since 2001, he has been with the Department of Mechanical and Environmental Informatics, Graduate School of Information Science and Engineering, Tokyo Institute of Technology, where he is currently a Professor. His research interests include control of nonlinear systems, and analysis and control of hybrid systems. He is an Associate Editor of *Automatica* and an Associate Editor of *SICE Journal of Control, Measurement, and System Integration*. He is a Member of the IEEE, SICE, ISCIE, IEICE and RSJ.