Ego-motion Noise Cancellation of a Robot using Missing Feature Masks

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

2011

Preprint:

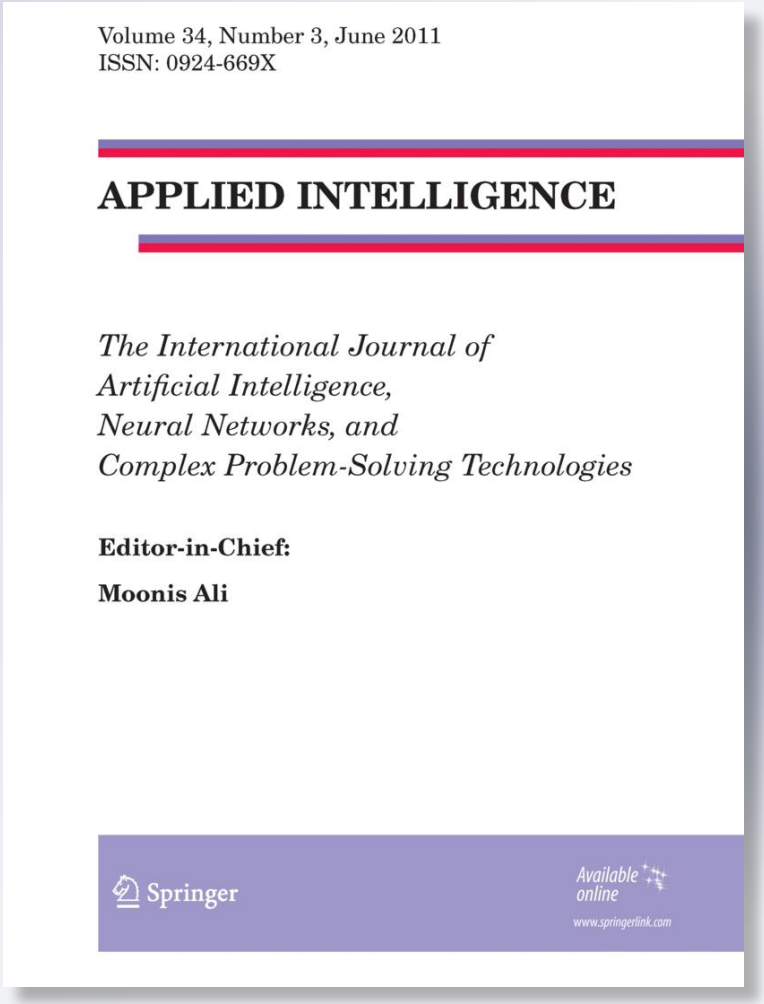
This is an accepted article published in Applied Intelligence. The final authenticated version is available online at: https://doi.org/[DOI not available]

Ego noise cancellation of a robot using missing feature masks

Applied Intelligence
The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies

ISSN 0924-669X
Volume 34
Number 3

Appl Intell (2011) 34:360-371
DOI 10.1007/
s10489-011-0285-0



Springer

Appl Intell (2011) 34: 360–371
DOI 10.1007/s10489-011-0285-0

Ego noise cancellation of a robot using missing feature masks

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

Published online: 29 March 2011

Abstract We describe an architecture that gives a robot the capability to recognize speech by cancelling ego noise, even while the robot is moving. The system consists of three blocks: (1) a multi-channel noise reduction block, comprising consequent stages of microphone-array-based sound localization, geometric source separation and post-filtering; (2) a single-channel noise reduction block utilizing template subtraction; and (3) an automatic speech recognition block. In this work, we specifically investigate a missing feature theory-based automatic speech recognition (MFT-ASR) approach in block (3). This approach makes use of spectro-temporal elements derived from (1) and (2) to measure the reliability of the acoustic features, and generates masks to filter unreliable acoustic features. We then evaluated this system on a robot using word correct rates. Furthermore, we present a detailed analysis of recognition accuracy to determine optimal parameters. Implementation of the proposed MFT-ASR approach resulted in significantly higher recognition performance than single or multi-channel noise reduction methods.

Keywords Ego noise · Noise reduction · Robot audition · Automatic speech recognition · Missing feature theory · Microphone array

Abbreviations

ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
BSS	Blind Source Separation
DoA	Direction of Arrival
GSS	Geometric Source Separation
HMM	Hidden Markov Model
MCRA	Minima Controlled Recursive Averaging
MFCC	Mel-Frequency Cepstral Coefficients
MFM	Missing Feature Mask
MFT	Missing Feature Theory
MMSE	Minimum Mean Square Estimation
MSLS	Mel-Scale Log Spectrum
MUSIC	MUltiple SIgnal Classification
NN	Nearest Neighbour
PF	Post-Filtering
SE	Speech Enhancement
SS	Spectral Subtraction
SSL	Sound Source Localization
SSS	Sound Source Separation
TS	Template Subtraction
WCR	Word Correct Rate
WF	Wiener Filtering

G. Ince (✉) · K. Nakadai · H. Tsujino
Honda Research Institute Japan Co., Ltd., 8-1 Honcho, Wako-shi, Saitama 351-0188, Japan
e-mail: gokhan.ince@jp.honda-ri.com

K. Nakadai
e-mail: nakadai@jp.honda-ri.com

H. Tsujino
e-mail: tsujino@jp.honda-ri.com

T. Rodemann
Honda Research Institute Europe GmbH, Carl-Legien Strasse 30, 63073 Offenbach, Germany
e-mail: tobias.rodemann@honda-ri.de

G. Ince · K. Nakadai · J. Imura
Dept. of Mechanical and Environmental Informatics, Tokyo Institute of Technology, 2-12-1-W8-1, O-okayama, Meguro-ku, Tokyo 152-8552, Japan

J. Imura
e-mail: imura@mei.titech.ac.jp

1 Introduction

Robots should be able to recognize and understand their auditory environments by using microphones to listen to their surrounding areas, an artificial listening capability known as "robot audition." However, the performance of some robot audition applications, such as Automatic Speech Recognition (ASR) and Sound Source Localization (SSL), is degraded drastically by background noise, reverberations, concurrent speakers and the robot's own noise. Signal quality and ASR accuracy may be improved by applying speech enhancement algorithms to the degraded speech. Since the short-term spectral characteristics of these noise signals differ substantially, various methods have been proposed to eliminate each individually. In general, robots with microphones are usually equipped with adaptive noise cancellation and acoustic echo cancellation methods for robust speech recognition in noisy and reverberant environments [1]. Microphone array-based multi-channel noise reduction methods [2] have shown especially good performance by forming and steering a beam in the direction of the desired sounds and attenuating interfering sound sources if the sound sources are directional, as in the case of multiple speakers. Although numerous signal processing techniques [2, 3] can deal with diffuse background noise, directional interfering noise and reverberations, the robot's own noise, also called *ego noise*, has not received as much attention.

Basically, the ego noise of a robot can be defined as the sum of fan noise, hardware noise and *ego-motion noise*. Fan noise comes from the fans that are located throughout the interior of the robot and help to dissipate the large amount of heat generated by the CPU, the power supply and other components, while hardware noise stems from the electrical circuits. The static (steady-state) fan noise and hardware noise can be removed easily by applying spectral filtering operations. In contrast, ego-motion noise is of special interest because it occurs only when the robot is performing an action using its motors. This special type of mechanical noise has so far either been ignored or circumvented due to its complex characteristics. The complexity is enhanced by the number of motors in action, making noise even more severe for a moving robot with a high degree of freedom. Nevertheless, mobility is a necessary condition for improving the perceptual capabilities of robots. Thus an autonomous robot with active perception may require a highly robust ability to suppress ego noise at any moment.

Ego-motion noise is more difficult to cope with than background or static fan noise, because it is non-stationary and, to a certain extent, similar to the signals of interest in terms of its directionality [4]. Therefore, conventional single-channel noise reduction methods, such as Spectral Subtraction (SS), Wiener Filtering (WF) and Minimum Mean Square Estimation (MMSE) [5–7], which incorporate estimates of the power spectrum of noise using the power spectrum of noisy speech [8–10], do not work well in practice. Treating ego-motion noise as a purely directional signal is also not a valid assumption, because ego-motion noise is also partially diffuse, due to the vibrations and reverberations inside the covers of the robot [4]. In practical terms, it is also not possible to localize each sound source (i.e., the noise signal emitted by each motor) at such short distances and cancel them spatially utilizing Sound Source Separation (SSS) techniques.

In this paper, we propose using template-based estimation, which is well-suited to capture the dynamic nature of the motion data represented by a sequence of observations. Based on these observations, it should be possible to associate a motion command or discrete time series data representing the motion (i.e., the angular status of each joint of the robot) with another series of discrete time data representing the total ego noise spectrum, thereby predicting an arbitrary sequence of associated data. This approach relies heavily on seamless synchronization between data on joint status (i.e., angular position, velocity and acceleration) and audio data. The high estimation quality achieved by this approach allows us to suppress noise accurately by applying a template-based spectral subtraction, also called Template Subtraction (TS). Furthermore, we incorporate Missing Feature Theory (MFT), which can be described as a filtering operation applied to missing or damaged acoustic features, to solve the ego noise problem of a robot at a higher level for an ASR application. To estimate the reliability of the features of speech, which are subject to residuals of motor noise after template subtraction, and to improve the performance of ASR, we propose to use MFT with a reliability estimation model based on ego-motion noise predictions. To generate suitable masks, we propose to integrate a multi-channel framework consisting of SSL, SSS, and Speech Enhancement (SE), in which the first two steps make use of the directionality of motor noise to cancel it and thus provide additional information about reliability. In contrast, the third step (SE) handles the diffuse portion of the ego-motion noise. In this respect, the main contribution of our work will be the incorporation of an original Missing Feature Mask (MFM) generation method based on the signals generated by two blocks (template subtraction & multi-channel noise reduction) that run in parallel. The mask relies on a measure of a frequency bin's quality calculated from the similarity of two totally different—yet complementary—approaches. We first suggest a binary mask, which uses either 0 or 1 to estimate the reliability of each acoustic feature. This method could be enhanced by using a soft mask, represented as continuous values between 0 and 1, which yields more detailed information about reliability. We demonstrate that these proposed methods achieve high noise elimination and thus ASR accuracy.

The rest of the article is organized as follows: In Sect. 2, we discuss related work on existing ego-motion noise suppression methods. Section 3 describes the proposed system and briefly summarizes the preprocessing stages, namely SSL, SSS, SE and template subtraction. Section 4 investigates a speech recognition system and computation of the missing feature masks in detail. Experiments and results are described in Sect. 5. The last section summarizes our conclusion and proposed future work.

2 Related work

Nakadai et al. [11] proposed a noise cancellation method that used two pairs of microphones. One pair in the inner part of the shielding body would record only internal motor noise and would help the sound localizer to determine whether each spectral sub-band is noisy or not, thus ignoring bands in which noise is dominant. In contrast to our approach, this technique does not suppress the noise. Several researchers addressed ego-motion noise problems by predicting and removing ego-motion noise using templates recorded in advance: For example, Ito et al. [12] used an Artificial Neural Network (ANN) to develop a frame-by-frame approach to cope with unstable walking noise. The trained network had to predict the noise spectrum resulting from the angular velocities of the joints of the robot. That study, however, concentrated on a small robot with limited degrees of freedom. Moreover, ANN has a slow training speed and online adaptation is difficult to achieve. In addition, this research was based mainly on estimations of templates for different motions, but did not focus on the possibility of quality improvement by utilizing spectral enhancement optimization factors. Ince et al. [13] proposed using parameterized template subtraction, incorporating tunable parameters to deal with variations in ego-motion noise, thus eliminating the training constraints of ANN. The accuracy of these templates was further enhanced by incorporating more information related to the joints, such as their angular acceleration. However, both of these methods [12, 13] suffer from the distorting effects of *musical noise* [6], which accompanies nonlinear single-channel based noise reduction techniques and reduces the intelligibility and quality of audio signals. In addition, this method, when utilized together with a nonlinear background noise prediction technique, e.g., Minima Controlled Recursive Averaging (MCRA) [5], results in a series of two consecutive nonlinear noise reduction operations. These operations produce even more musical noise, eventually damaging the acoustic features and reducing the recognition performance of ASR.

In the field of "Robot Audition", which pursues general understanding of sound, noise suppression is achieved mostly using sound source separation techniques with microphone arrays [14–16]. Ego-motion noise cannot be completely explained by a directional noise model, such as assumed for interfering speakers [14, 15] or by a diffuse background noise model [5]. Because the motors are located in the near field of the microphones, they produce sounds that have both diffuse and directional characteristics. In another related work, Even et al. [16] proposed using 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, and a delay-and-sum beam-former enhanced the refined speech. Although it improved speech recognition accuracy considerably, this method requires a body cover made of high-quality or thick material so that no external noise could be recorded by these additional sensors. In contrast, our method can be implemented on any mobile robot, with no physical constraints on its external shielding, and utilizes only existing microphones.

Research has also focused on specific conditions for near field sound sources. For example, Mizumachi et al. described a model utilizing line sound sources and spherical wave propagation in the near field, in contrast to conventional far field assumptions such as plane wave propagation and point-shaped sound sources [17]. Zheng et al. proposed a spherically isotropic noise model for near field objects, achieving greater suppression of reverberations and reduction in beam-pattern variations for broadband signals, similar to our motor noise signals [18]. The proposed models, however, are computationally expensive, can deal with only a single sound source, and more importantly, are designed for stationary sound sources. In a standard task utilizing moving robots where the acoustic conditions of the noise, such as the power and frequencies of the motor noise spectrum as well the number of active motors, can dynamically change over time, the performance of these models can decrease drastically.

From the perspective of signal processing, unreliable features of speech have been shown to degrade recognition performance severely [19]. MFT has already found useful applications, such as recognition of speech corrupted by music and several types of noise (refer to [19] for a comprehensive study). For a simultaneous speech recognition task of several speakers, Yamamoto et al. [14] and Takahashi et al. [20] have proposed a model for mask generation based on the disturbing effect of leakage noise over speech, because an imperfect source separation causes distortion. Their model, however, was unable to deal with ego-motion noise. Nishimura et al. [21] estimated the ego noise of distinct gestures and motions of the robot. Using motion commands, the pre-recorded correct noise template matching a recent motion could be selected from the template database and the acoustic features of the aligned template could be used for MFT weight calculation. Since their mask model

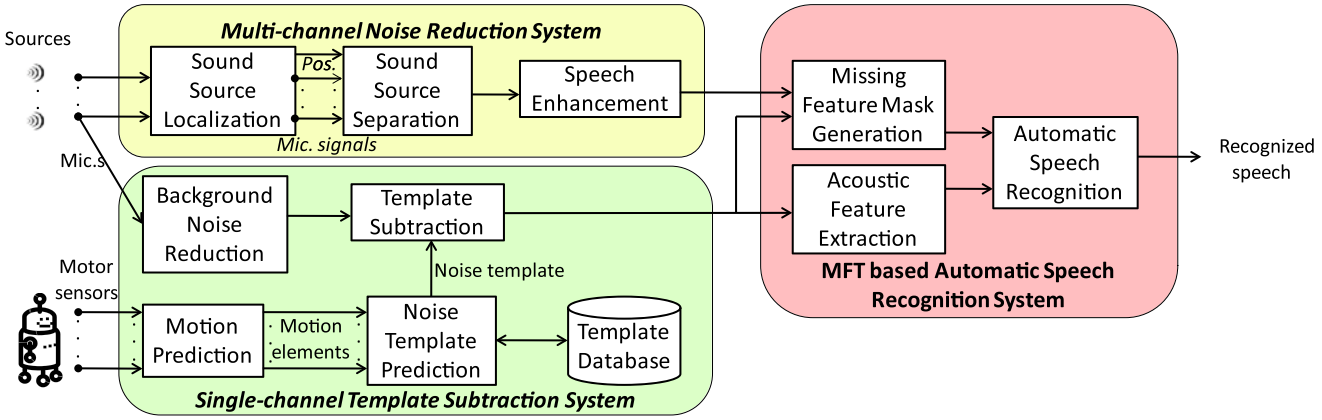


Fig. 1 Proposed noise cancellation system

was based on a simple energy threshold, it was not feasible for use in a real-world scenario, where the Signal-to-Noise Ratio (SNR) of speech can change depending on the loudness and distance of the speaker from the robot. They also utilized blockwise templates, which could not cope with dynamic changes in motion trajectories over time. Our approach overcomes the former problem by introducing a mask generation method that operates on speech signals refined from all types of noise, and overcomes the latter problem by using a parameterized template prediction method, as in [13].

3 System overview

As sensors we used an array of multiple omnidirectional microphones mounted around the head of the robot. The overall architecture of the proposed noise reduction system is shown in Fig. 1. The first block of our processing chain, composed of elements for performing SSL, extracts the location of the most dominant sources of noise in the environment. The locations of these sources can be estimated using a linear separation algorithm called Geometric Source Separation (GSS) [14], which can be considered a hybrid algorithm of Blind Source Separation (BSS) [22] and beam-forming. SSS is followed by a speech enhancement step, called multichannel Post-Filtering (PF). This module attenuates stationary noise, i.e., background noise, and non-stationary noise that arises from the leakage energy between the output channels of the previous separation stage for each individual sound source. These three modules constitute the **multi-channel noise reduction** block [4], while a second block performs **template subtraction** [13]. These two modules together are responsible for the *audio features* of speech recognition and *spectrograms* that will be further processed in the MFM generation stage. Finally, a new third block, **MFT-based speech recognition**, designed to achieve a more robust ASR, uses both the features and spectrograms created in the pre-processing stages to extract the most suitable features. This part will be discussed in Sect. 4 in detail.

3.1 Multi-channel noise reduction system

To estimate the Direction of Arrival (DoA) of each sound source, we used a popular adaptive beamforming algorithm called MUltiple SIgnal Classification (MUSIC) [23]. This algorithm identifies the DoA by performing eigenvalue decomposition on the correlation matrix of the noisy signal, by separating subspaces of undesired interfering sources from sound sources of interest, and by identifying the peaks in the spatial spectrum. A consequent source tracker system performs temporal integration in a given time window.

Geometric Source Separation [22], later on extended to be an adaptive algorithm that can process input data incrementally [15], explicitly makes use of the locations of sources. To estimate the separation matrix properly, GSS introduces cost functions that must be minimized in an iterative way (see [15] for details). Moreover, we used adaptive step-size control, which results in rapid convergence of the separation matrix [24]. Our GSS implementation also exploited a method called Optima Controlled Recursive Averaging, which controls the window size adaptively causing a smoother convergence and thus better separation results [25].

After the separation process, a multi-channel post filtering operation was utilized to further enhance the sounds. This module is based on the optimal estimator proposed by Ephraim and Malah [26]. Since their method takes temporal and spectral continuities into consideration, it generates less distortion than conventional spectral subtraction-based noise reduction methods. To further extend this concept, we applied a multi-channel post filter [15], which can cope with non-stationary interferences as well as stationary noises. This module treats the transient components in the

spectrum as if they are caused by leakage energies that may occasionally arise due to poor separation performance. For this purpose, noise variances of both stationary noise and source leakage are predicted, the former using the MCRA method [6], and the latter using the algorithm proposed in [15]. The noise suppression rule further involves speech presence probability calculations [27] and is based on minimum mean-square error estimation of the spectral amplitude [26].

3.2 Single-channel template subtraction system

The underlying motivation for using templates for noise reduction is that the duration of the motor noise signals does not change for the same motions by more than a few samples when the motion is performed again and the magnitude of the noise signal does not deviate much from the mean magnitude of a set of same motions. Thus, we made the following assumptions:

- Current motor noise depends on the position, velocity and acceleration of that specific motor.
- Similar combinations of joint status will result in similar motor noise spectral vectors at any time point.
- The superposition of each single joint motor noise at any time point is equal to the whole body noise at that time.

We defined a *template* as the representation of an actual segment of noise. A *blockwise template* fundamentally consists of two parts: 1) Label: Motion label (e.g., "wave right hand" or "turn head from 0° to 40°"), and 2) Data: Whole ego noise spectrum recorded during the motion (i.e., spectral matrix). A conventional blockwise template subtraction first estimates the ego noise spectrum using only motion commands as representative labels for the corresponding ego noise spectrum and then removes it from the noisy spectrum to obtain a clean speech spectrum. Basically, a query in the motion command labels from a database enables the selection of the appropriate ego noise template, recorded from the onset until the offset time of motor noise. This method, however, has several shortcomings; e.g., it could be performed properly only after the detection of the exact starting moment of the template, which is a very hard task to achieve. Otherwise, it suffers from misalignments of the templates in time. Furthermore, this method requires a huge amount of data for each possible motion. Considering the impossibility of collecting and producing templates for each joint of different combinations of origin, target, position, velocity and acceleration parameters, this approach is simply not feasible in a realistic scenario. Finally, this rather primitive representation of motion labels cannot deal with deviations in motion trajectories.

To overcome these deficits, we developed a technique that parameterizes a discrete audio segment under consideration using motor status, obtaining a spectral vector that represents the ego noise at that instant of time. This *parameterized template* has a different structure than the blockwise template: 1) Label: Instantaneous joint status of the robot (i.e., feature vector), and 2) Data: Instantaneous ego noise spectrum of one frame (i.e., spectral vector). To implement so-called parameterized template subtraction [13], we need 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 by exerting spectral enhancement parameters, as shown in Sect. 3.2.3.

Before explaining the details of parameterized template subtraction, we first define an input signal $y(t)$ at time t, which can be expressed as

$$y(t) = x(t) + d(t), \tag{1}$$

where $x(t)$ is a target signal and $d(t)$ is the distortion, i.e., noise signal. Noise estimation and reduction algorithms operate in the time-frequency (spectrogram) domain. The complex input spectrum $Y(\omega, k)$ of discrete frequency bin ω and time frame k is obtained from

$$Y(\omega, k) = \sum_{t=0}^{t=F-1} y(t + kM)w(t) \exp\{j(2\pi/F)t\omega\}, \tag{2}$$

where F is the window length, M is the shift length and $w(t)$ is the window function. Finally, the spectrum of the observed signal $Y(\omega, k)$ can be given as:

$$Y(\omega, k) = X(\omega, k) + D(\omega, k). \tag{3}$$

3.2.1 Generation of the template database

To generate the template database, joint status information provided by the sensors on the motors will be utilized. During the motion of the robot, the actual position (θ) information for each motor is gathered regularly. Using the difference between consecutive sensor outputs, velocity ($\dot{\theta}$) and acceleration ($\ddot{\theta}$) can be calculated. If J joints are active, $3J$ attributes are generated. Each feature is normalized to $[-1 \quad 1]$, so that all features have the same contribution to the prediction. The resulting feature vector has the form, $\mathfrak{F}(k) = [\theta_1(k), \dot{\theta}_1(k), \ddot{\theta}_1(k), \ldots, \theta_J(k), \dot{\theta}_J(k), \ddot{\theta}_J(k)]$. At the same time, motor noise is recorded and background, static fan and hardware noise is removed from the overall noise recordings. The spectrum of the motor noise ($D(k) = [D(1, k), D(2, k), \ldots, D(F, k)]$, 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 corresponding templates can be generated when their

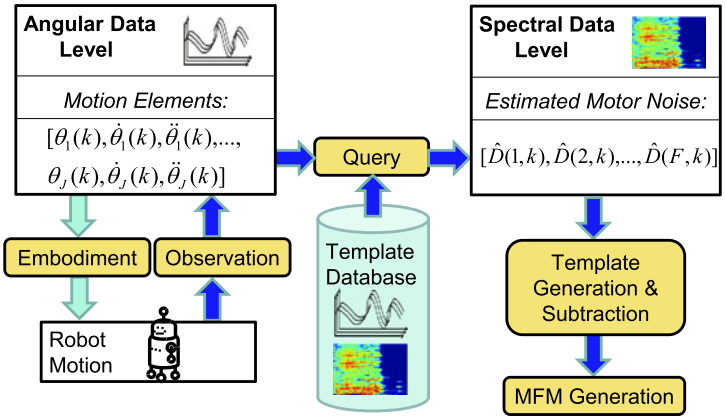


Fig. 2 Parameterized template prediction method and its applications for ego-motion noise robust speech recognition

time tags match. Each parameterized template is in the format of a data block consisting of two concatenated vectors, $[\mathfrak{F}(k), D(k)]$. Finally, a large noise template database, consisting of short noise templates for many joint configurations, can be created.

3.2.2 Motor noise prediction

The prediction phase starts with a search of the database for the best matching template of motor noise at that time (Fig. 2). Finding the correct template involves a search of all the templates in the database for most similar joint configuration. We utilized a 1-Nearest Neighbor (1-NN) search to accomplish this task. The spectral vector, $\hat{D}(k) = [\hat{D}(1,k), \hat{D}(2,k), \ldots, \hat{D}(F,k)]$, associated with the point in the database with the shortest distance to the query point was selected as the template. This prediction process can be applied to every frame. In that sense, a template for a single arbitrary motion of an arbitrary duration can be regarded as the concatenation of smaller templates predicted according to the above-mentioned approach on a frame-by-frame basis.

3.2.3 Template subtraction

The power spectrum of the useful signal can be obtained by applying:

$$X_r(\omega,k) = Y(\omega,k) - \hat{D}(\omega,k), \tag{4}$$

where $\hat{D}(\omega,k)$ denotes the estimated noise template and $X_r(\omega,k)$ is the signal comprising the useful sound and residual motor noise, which results from the deviation of the original motor noise $D(\omega,k)$ from the predicted motor noise. To compensate for this error, we used a spectral subtraction approach that utilized an *overestimation factor*, α, and a *spectral floor*, β. α, also called an *aggressiveness factor*, thus allowing 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 non-linear mapping of the negative or small-valued spectral estimates. This produces a metallic noise that sounds like the sum of tone generators with random fundamental frequencies, which are turned on and off constantly [5]. β reduces the effect of the sharp valleys and peaks in the spectrum, which are caused by the smaller attenuations of these compared with neighboring frequencies due to the random fluctuations in the estimations of magnitude.

A noise reduction process can be divided into two consequent processes: gain calculation and spectral filtering. The gains can be calculated using the formula of magnitude spectral subtraction [5]:

$$\hat{H}_{SS}(\omega,k) = max\left(1 - \alpha\frac{|\hat{D}(\omega,k)|}{|Y(\omega,k)|}, \beta\right), \tag{5}$$

where $\hat{H}_{SS}(\omega,k)$ represents the gain coefficient and *max* is the maximum value calculation. A spectral filtering operation of the signal $Y(\omega,k)$ with this coefficient finalizes the *template subtraction* as in (6):

$$\hat{X}(\omega,k) = \hat{H}_{SS}(\omega,k) \cdot Y(\omega,k). \tag{6}$$

Unlike previous methods [12, 21], our prediction, generation and subtraction methods did not require any starting or ending signals, indicating that no abrupt blockwise templates are applied to the noisy signals discontinuously. Our methods continuously process the data, 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. We conducted a training session of uninterrupted sound recording for a single continuous motion sequence consisting of hundreds of individual motions with short (<1 sec.) pauses between each motion.

4 MFT-based automatic speech recognition system

Different strategies, which make use of a confidence-based weighting of the time-frequency representation of audio signals, can enhance the quality of speech. Missing Feature Theory is a promising approach that basically applies a mask to decrease the contribution of unreliable parts of distorted speech [19]. Retaining the reliable parameters that are essential for speech recognition results in a substantial increase in recognition accuracy [14, 20]. In this section, we will discuss the basic steps of such an ASR system and how this approach can be adapted to the ego noise problem by presenting a robust mask design method for estimating the reliability of speech based on current motor noise.

4.1 Acoustic feature extraction

Acoustic features are extracted from the refined spectra, the final products of the noise reduction stages (see Fig. 1). An additive white noise step applied after post-filtering improves speech recognition results, due to the generation of an artificial spectral floor in the background of a speech signal and the blurring of musical noise distortions. Because we did not want the distortions to spread to all coefficients of the cepstrum, we avoided using Mel-Frequency Cepstral Coefficients (MFCC). Rather, we used the Mel-Scale Log Spectrum (MSLS), whose detailed calculation method has been described in [28]. Moreover, linear regression of each spectral coefficient is represented as a *delta* feature and used to enhance the quality of acoustic features. Spectral mean normalization improved the noise robustness of MSLS features by subtracting the average of the features in the previous 5 seconds from the current features.

4.2 MFM generation

The reliability of features is computed for each frame and for each mel-frequency band. Masks composed of continuous values between 0 and 1 are called *soft masks*, whereas masks composed of only discrete values, either 0 or 1, are called *hard masks*. We assessed the performance of both methods for this particular type of ego noise problem. We start by explaining some basic ego-motion noise suppression capabilities of the preprocessing stages of our proposed system. Then, we show how to derive the masks in detail.

GSS lacks the ability to identify and suppress motor noise originating from the same direction as the speaker, because it regards the noise as part of the speech. Moreover, if the position of the noise source is not detected precisely, GSS cannot separate the sound in the spatial domain. Thus, small amounts of motor noise can spread to the separated sound sources. However, multi-channel noise suppression systems work very well for weaker motion noise, such as arm or leg motions, when compared with head motion noise, as we demonstrated [4]. In addition, our system was optimally designed for "simultaneous multiple speakers" scenarios with background noise and demonstrated a very good performance when no motor noise was present.

In contrast, template subtraction makes no assumptions about the directivity or diffuseness of the sound source and can match a pre-recorded template of the motor noise at any moment. The drawback of this approach, however, is that due to its not being stationary, the characteristics of predicted and actual noise may differ to some extent.

As described, the two approaches have distinct strengths and weaknesses and thus may be used in a complementary fashion. A speech feature is considered unreliable if the difference between the energies of refined speech signals generated by multi-channel and single-channel noise reduction systems is above a threshold T. The masks are computed for each frame, k, and for each frequency band, f. First, a continuous mask is calculated as:

$$m(f,k) = \frac{|\hat{S}_m(f,k) - \hat{S}_s(f,k)|}{\hat{S}_m(f,k) + \hat{S}_s(f,k)}, \tag{7}$$

where $\hat{S}_m(f,k)$ and $\hat{S}_s(f,k)$ are the estimated energy of the refined speech signals ($|\hat{X}(\omega,k)|^2$), following multi-channel noise reduction and single-channel template subtraction, respectively. Both signals are computed using a mel-scale filterbank. The numerator represents the deviation of the two outputs, a measure of their uncertainty or unreliability. The denominator, however, is a scaling constant and is the average of the two estimated signals. (To simplify the equation, we removed the scalar values from the denominator, so that $m(f,k)$ can take on values between 0 and 1.) Thereby, the reliability can be defined by $\frac{1}{m(f,k)}$, which means the smaller the $m(f,k)$, the higher the reliability and vice versa. Depending on the type of mask (hard or soft) used in the MFT-ASR, (8) or (9) is selected.

1. For hard (binary) masks:

$$M(f,k) = \begin{cases} 1, & \text{if } m(f,k) < T, \\ 0, & \text{if } m(f,k) \geq T. \end{cases} \tag{8}$$

2. For soft masks [20]:

$$M(f,k) = \begin{cases} \dfrac{1}{1+\exp(-\sigma(m(f,k)-T))}, & \text{if } m(f,k) < T, \\ 0, & \text{if } m(f,k) \geq T, \end{cases} \tag{9}$$

where σ is the tilt value of a sigmoid weighting function.

4.3 MFT-ASR

Missing Feature Theory Based Automatic Speech Recognition (MFT-ASR) is a Hidden Markov Model (HMM) based speech recognition technique [19]. If $M(i)$ is the MFM vector generated as in Sect. 4.2 for the i-th acoustic feature, the output probability can be expressed as:

$$b_j(x) = \sum_{l=1}^{L} P(l|S_j) \exp\left\{ \sum_{i=1}^{I} M(i) \log f(x(i)|l, S_j) \right\}, \tag{10}$$

where $b_j(x)$ is the output probability of the j-th state, $x(i)$ is an acoustic feature vector, I is the size of the acoustic feature vector, $P(\cdot)$ is the probability operator and S_j is the j-th state. Density in each state S_j is modeled using mixtures of L Gaussians with diagonal-only covariance. When all mask values are set to 1, (10) is identical to the output probability calculation of a conventional ASR.

5 Results

In this section, we compare results of the proposed system with those of existing approaches against ego-motion noise, after describing the experimental settings and environment. We assessed the performance of pre-processing based ASR, hard and soft mask based ASR, and the influence of selected parameters for template subtraction and MFT-ASR blocks on performance.

5.1 Experimental settings

To evaluate the performance of the proposed techniques, we used a humanoid robot developed by Honda. This robot, which has many degrees of freedom, was equipped with an 8-ch microphone array on top of its head. We used 2 motors for head motion and 4 motors for the motion of each arm, resulting in 10 degrees of freedom. Sensors recorded the angle of each joint every 5 ms and the length of each audio frames was 10 ms. We used a constant $\alpha = 1$ and varying β values as template subtraction parameters, because we had observed that higher values of α damage speech and an increase in β improved ASR accuracy considerably compared with $\beta = 0$. (For detailed evaluations of the parameters α and β, their effects on ASR accuracy, signal quality and noise suppression rates, see [13].)

We recorded random motions performed by the given set of limbs by storing a training database of 30 minutes' duration. In a separate session, we recorded a test database 10 minutes long for evaluation. Because the noise recordings were longer than the utterances used in isolated word recognition, we selected those segments, in which all joints contributed to the 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, SNR_{seg}, which estimates the SNR-level within each segment and averages them over the whole signal, providing a better representation of the energy distribution of speech and noise within the time interval of interest.

$$\mathrm{SNR}_{seg} = \frac{1}{J} \sum_{j=1}^{J} 10 \log_{10} \left(\frac{\sum_t x_j^2(t)}{\sum_t d_j^2(t)} \right), \tag{11}$$

where J is the number of segments with speech activity, and $x(t)$ and $d(t)$ are the t-th discrete speech and noise sample respectively. The noise signal, consisting of whole ego noise and environmental background noise, was mixed with clean speech utterances used in a typical human-robot interaction dialog. This Japanese word dataset contains 236 words spoken by 4 female and 4 male speakers. Acoustic models are trained with Japanese Newspaper Article Sentences (JNAS) corpus, 60-hours of speech data spoken by 306 people (153 males and 153 females), making speech recognition a speaker-open and word-open test. We used 13 static MSLS, 13 delta MSLS and 1 delta power for an acoustic feature. Speech recognition results were reported as average Word Correct Rates (WCR, defined as the number of correctly recognized words from the test set divided by the number of all instances in the test set). *WCR improvement* was calculated by subtracting two WCRs obtained from the experiments with two different parameter sets or two different methods and represented as "points." The position of the speaker was fixed at 0° throughout the experiments. The recording environment consisted of a room 4.0 m × 7.0 m × 3.0 m in size with a reverberation time (RT_{20}) of 0.2 sec.

The implementation runs on HARK, an open-sourced software program for robot audition [29]. Although the position of the original sound source was provided in advance to avoid mis-recognition due to localization errors, we did not fix the ego-noise direction of the robot. In this experiment, the SSL module predicted it automatically.

5.2 Spectrograms and masks

Figure 3 shows a general overview of the effect of each processing stage until the masks are generated. Figure 3c represents a dense mixture of speech (Fig. 3a) and motor noise (Fig. 3b) with an SNR of −5 dB. GSS+PF in Fig. 3g reduced only a minor part of the motor noise while sustaining the speech. In contrast, template subtraction (Fig. 3h) reduced the motor noise aggressively while distorting some parts of the speech. The hard mask (Fig. 3i) provides a filter that eliminates unreliable and still noisy parts of the speech ($T = 0.5$). The soft mask (Fig. 3j) provides more detailed information about the degree of reliability of each feature

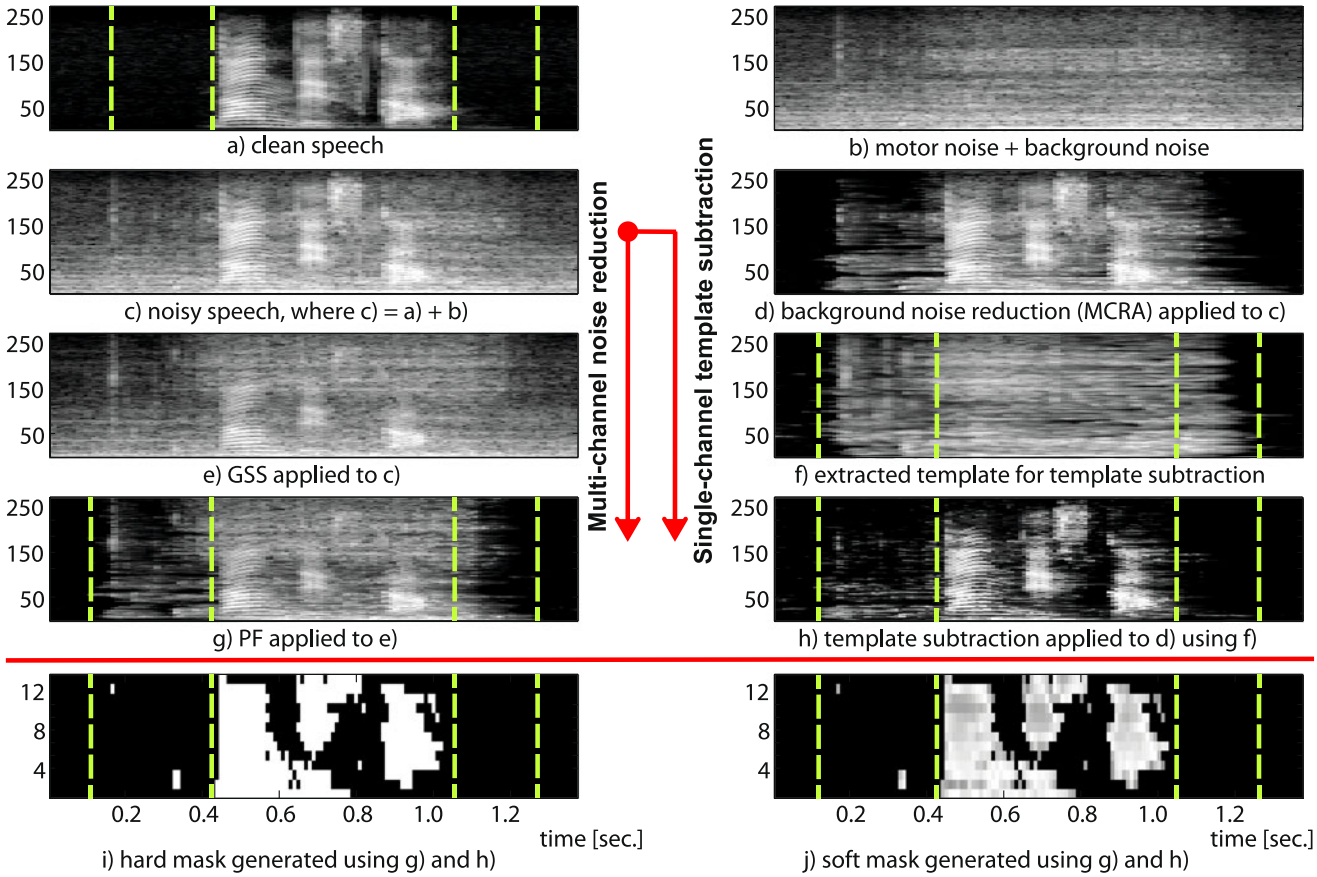


Fig. 3 Spectra of speech signal (utterance: "Nan desu ka?" (What is this?)), noisy speech signals, refined speech signals and corresponding masks. In (**a**)–(**h**), the y-axis represents 256 frequency bins between 0 and 8 kHz and in (**i**)–(**j**) the y-axis represents 13 static MSLS features. The x-axis in all panels is the index of frames

so that the noise-free features are weighted more than the noise-containing parts of the MFT-ASR ($\{T, \sigma\} = \{0.5, 5\}$). Furthermore, we found that features between times 0.10–0.42 sec. and 1.07–1.27 sec., which were composed basically of motor noise alone, were given zero weights in the masks, except for a few mis-detections. The dotted yellow lines in the panels of Fig. 3 indicate the borders of these regions, with speech features located between the 0.42–1.07 sec. Within this time interval, these masks were able to detect even those speech features that were contaminated by motor noise residuals and set either zero or low weights.

5.3 ASR accuracy using MFMs

We compared our MFT-based noise elimination approach with the single-channel noise suppression (TS) and multi-channel (GSS+PF) noise suppression techniques. The results were evaluated using an acoustic model trained with MCRA-applied speech data, except that, for the GSS+PF method, we used a matched acoustic model of this particular condition. In preliminary tests, we found that the feature set derived at the output of template subtraction achieved a greater accuracy by 10–20 points in WCR, compared with the features after multi-channel noise reduction. We therefore concluded that the former feature type is more suitable for an MFT-ASR. Single-channel results were used as a baseline for comparing all ego-motion noise reduction methods. Figure 4 illustrates the ASR accuracies for all methods under consideration. MFT-ASR outperformed both single (TS) and multi-channel (GSS+PF) noise reduction methods. We also evaluated MFMs for three heuristically selected threshold parameters, $T = \{0.25, 0.5, 0.75\}$, with the outcomes presented in Fig. 5. If $T < 0.5$, ASR was not improved because essential features belonging to speech were discarded, resulting in a deterioration of WCRs. In contrast, higher thresholds improved the outcomes significantly.

In our second set of experiments, we compared the results of hard masking with an optimal threshold ($T = 0.75$) obtained during our first set of experiments, with the results of soft masking for the parameter set $\sigma = \{5, 10, 50\}$. All three examples with these parameters yielded similar WCR improvements. Outside this range, however, the results become sensitive to σ and eventually deteriorated. Therefore, we will present only the results for $\sigma = 5$. We also assessed

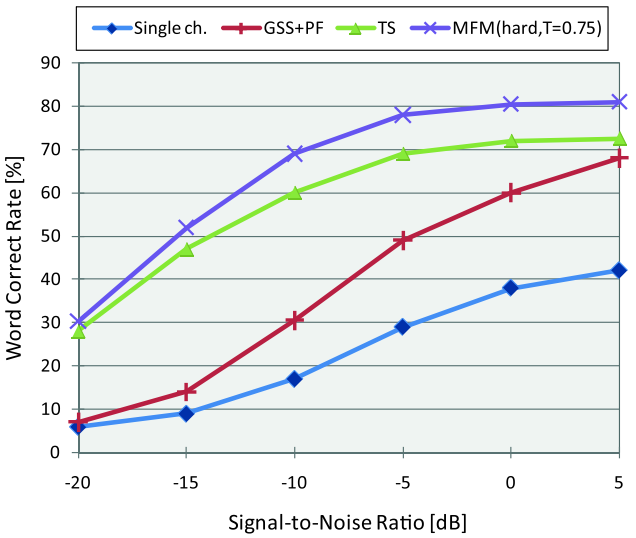


Fig. 4 Speech recognition performance of different processing stages

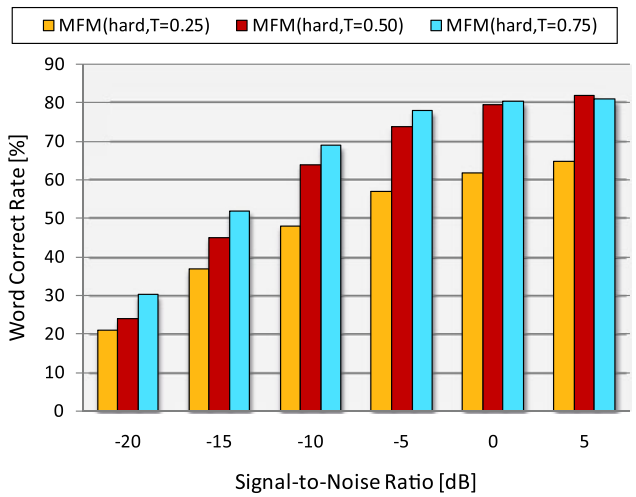


Fig. 5 Speech recognition performance of different MFM settings

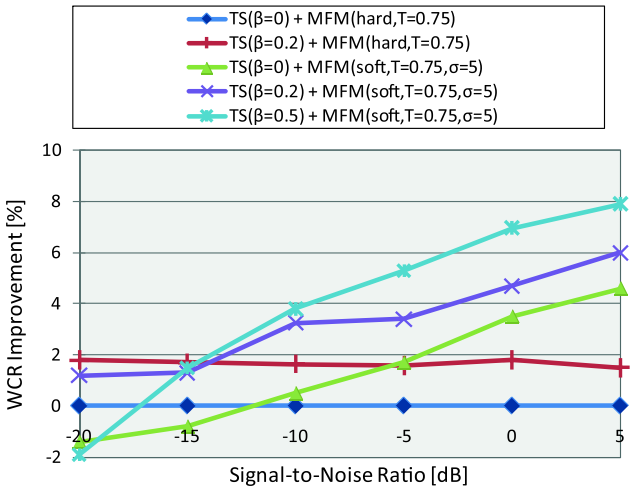


Fig. 6 Speech recognition performance of soft mask–hard mask comparisons for given parameters

Table 1 Recognition accuracies (% WCR) of all methods utilized in this study with realistic SNRs during a robot-human interaction (SNR = {−5, 0} dB)

SNR	Sing. ch.	GSS + PF	TS	TS + MFM ($T = 0.75$)		
				Hard	Soft	
					$\beta = 0$	$\beta = 0.5$
−5 dB	29	49	69	78	80	**84**
0 dB	38	60	72	80	85	**87**

the effect of decreasing the aggressiveness level of the template subtraction, by leaving an artificial floor on the bottom of the spectra. Thus, in our first set of experiments, the parameter called *spectral floor* (β, where $0 \leq \beta \leq 1$) [13] was set to zero. We assessed the results for $\beta = \{0, 0.2, 0.5\}$ in the framework of soft-hard mask comparison in Fig. 6 by determining the improvement in WCR relative to that obtained for the hard mask at $\beta = 0$ and $T = 0.75$. Increasing β resulted in considerable improvements in the WCRs, indicating that a tradeoff between "noise reduction level" and "signal distortion" contributed substantially to the quality of the mask. We found that soft masks improved the WCRs even further by up to 8 points compared with hard masks. This reduction was due to the improved probabilistic representation of the reliability of each feature. Optimal results were obtained when we used a soft mask with the parameter set: $\{T, \sigma, \beta\} = \{0.75, 5, 0.5\}$.

While the masks eliminated unreliable speech features contaminated with motor noise, they also compensated for the erroneous effects of voice activity detection due to additive motor noise containing a large portion of energy. These masks prevented the false identification of motor noise as speech, when speech had not yet started, or had been completed. Table 1 shows the average WCRs extracted from the results in Figs. 4 and 6. This table helps in visualizing the simulated results in a real-world scenario with a robot, where SNRs usually vary by [−5∼0] dB for head and arm motion noises depending on the optimal distance and loudness of the speaker. The gain achieved by applying soft masking was 15 points greater than that of the single-channel template subtraction method.

6 Conclusion

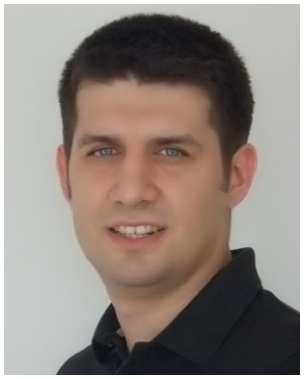
We have presented a method for eliminating ego noise from speech signals. This system utilizes (1) a multi-channel noise reduction stage, (2) a template subtraction stage, and

finally (3) a masking stage to improve speech recognition accuracy. We used an MFM model, which is based on the similarity of measurements of ego noise estimations obtained from (1) and (2). We validated the applicability of our method by evaluating its performance at different settings for hard and soft MFMs. Our method demonstrated significant WCR improvements with hard masking (49 points relative to single-channel recognition) and soft masking (up to 55 points).

In future, we intend to determine an optimized parameter set for template subtraction for this specific MFT-ASR task in a wider range. The next step is an evaluation online and in a real environment, involving the recognition of speech by several speakers simultaneously while the robot is performing some motions. We also plan to conduct experiments on a robotic system, which does not provide perfectly synchronous data streams of motion and audio, to evaluate the limitations of our method. It is our intention to extend the method so that it is able to cope with incoming data streams, which are asynchronous to some extent.

References

1. Sato M, Sugiyama A, Ohnaka S (2004) An adaptive noise canceller with low signal-distortion based on variable stepsize subfilters for human-robot communication. IEICE Trans Fundam Electron Commun Comput Sci E88-A(8):2055–2061
2. Brandstein M, Ward D (2001) Microphone arrays: signal processing techniques and applications. Springer, Berlin
3. Benesty J, Sondhi MM, Huang Y (2008) Springer handbook of speech processing. Springer, Berlin
4. Ince G, Nakadai K, Rodemann T, Hasegawa Y, Tsujino H, Imura J (2010) A hybrid framework for ego noise cancellation of a robot. In: Proceedings of the IEEE international conference on robotics and automation (ICRA), pp 3623–3628
5. Boll S (1979) Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans Acoust Speech Signal Process 27(2):113–120
6. Cohen I (2002) Noise estimation by minima controlled recursive averaging for robust speech enhancement. IEEE Signal Process Lett 9(1):12–15
7. Deller J (2000) Discrete-time processing of speech signals. IEEE Press, New York
8. Martin R (1994) Spectral subtraction based on minimum statistics. In: Proceedings European signal processing, pp 1182–1185
9. Cohen I, Berdugo B (2001) Speech enhancement for non-stationary noise environments. Signal Process 81:2403–2481
10. Nakajima H, Ince G, Nakadai K, Hasegawa Y (2010) An easily-configurable robot audition system using histogram-based recursive level estimation. In: Proceedings of the IEEE/RSJ international conference on robots and intelligent systems (IROS), pp 958–963
11. Nakadai K, Okuno HG, Kitano H (2000) Humanoid active audition system improved by the cover acoustics. In: PRICAI 2000 topics in artificial intelligence (sixth pacific rim international conference on artificial intelligence). Springer lecture notes in artificial intelligence, vol. 1886. Springer, Berlin, pp 544–554
12. Ito A, Kanayama T, Suzuki M, Makino S (2005) Internal noise suppression for speech recognition by small robots. In: Proceedings of the interspeech, pp 2685–2688
13. Ince G, Nakadai K, Rodemann T, Hasegawa Y, Tsujino H, Imura J (2009) Ego noise suppression of a robot using template subtraction. In: Proceedings of the IEEE/RSJ international conference on robots and intelligent systems (IROS), pp 199–204
14. Yamamoto S, Nakadai K, Nakano M, Tsujino H, Valin J-M, Komatani K, Ogata T, Okuno HG (2006) Real-time robot audition system that recognizes simultaneous speech in the real world. In: Proceedings of the IEEE/RSJ international conference on robots and intelligent systems (IROS), pp 5333–5338
15. Valin J-M, Rouat J, Michaud F (2004) Enhanced robot audition based on microphone array source separation with post-filter. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 2123–2128
16. Even J, Sawada H, Saruwatari H, Shikano K, Takatani T (2009) Semi-blind suppression of internal noise for hands-free robot spoken dialog system. In: Proceedings of the IEEE/RSJ international conference on robots and intelligent systems (IROS), pp 659–663
17. Mizumachi M, Nakamura S (2004) Passive subtractive beamformer for near-field sound sources. In: Proceedings of the IEEE sensor array and multichannel signal processing workshop, pp 74–78
18. Zheng YR, Goubran RA, El-Tanany M (2003) A nested sensor array focusing on near field targets. In: Proceedings of the IEEE sensors, vol 2, pp 843–848
19. Raj B, Stern RM (2005) Missing-feature approaches in speech recognition. IEEE Signal Process Mag 22:101–116
20. Takahashi T, Yamamoto S, Nakadai K, Komatani K, Ogata T, Okuno HG (2008) Soft missing-feature mask generation for simultaneous speech recognition system in robots. In: Proceedings of the interspeech, pp 992–997
21. Nishimura Y, Ishizuka M, Nakadai K, Nakano M, Tsujino H (2006) Speech recognition for a robot under its motor noises by selective application of missing feature theory and MLLR. In: Proceedings of the IEEE-RAS international conference on humanoid robots, pp 26–33
22. Parra LC, Alvino CV (2002) Geometric source separation: merging convolutive source separation with geometric beamforming. IEEE Trans Speech Audio Process 10(6):352–362
23. Schmidt R (1986) Multiple emitter location and signal parameter estimation. IEEE Trans Antennas Propag 34(3):276–280
24. Nakajima H, Nakadai K, Hasegawa Y, Tsujino H (2008) Adaptive step-size parameter control for real-world blind source separation. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 149–152
25. Nakadai K, Nakajima H, Hasegawa Y, Tsujino H (2009) Sound source separation of moving speakers for robot audition. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 3685–3688
26. Ephraim Y, Malah D (1984) Speech enhancement using minimum mean-square error short-time spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process 32(6):1109–1121
27. Cohen I, Berdugo B (2002) Microphone array post-filtering for non-stationary noise suppression. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 901–904
28. Nishimura Y, Shinozaki T, Iwano K, Furui S (2004) Noise-robust speech recognition using multi-band spectral features. In: Proceedings of the 148th acoustical society of America meetings 1aSC7
29. Nakadai K, Takahashi T, Okuno H, Nakajima H, Hasegawa Y, Tsujino H (2010) Design and implementation of robot audition system "HARK"—open source software for listening to three simultaneous speakers. Adv Robot 24:739–761



Gökhan Ince received the B.S. degree in electrical engineering from Istanbul Technical University, Turkey, in 2004, the M.S. degree in information engineering in 2007 from Darmstadt University of Technology, Germany. He is currently pursuing a Ph.D. 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 is 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 IEEE, RAS, ISAI and ISCA.

Kazuhiro Nakadai received B.E. in electrical engineering in 1993, M.E. in information engineering in 1995, and Ph.D. in electrical engineering in 2003 from Tokyo University. He had been worked with Nippon Telegraph and Telephone and NTT Comware Corporation for four years as a system engineer from 1995 to 1999. He was working with Kitano Symbiotic Systems Project, ERATO, Japan Science and Technology Agency (JST) as a researcher from 1999 to 2003. He is currently principal researcher for Honda Research Institute Japan, Co., Ltd. From 2006 to 2010, he was concurrently Visiting Associate Professor at Tokyo Institute of Technology, and he is Visiting Professor at Tokyo Institute of Technology since 2011. He also has another position of Visiting Professor at Waseda University from 2011. His research interests include AI, robotics, signal processing, computational auditory scene analysis, multi-modal integration and robot audition. He is a member of 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 Ph.D. degree from the Technische Universität Bielefeld, Germany in 2003. Since 1998 he is working at the Honda Research Institute Europe, Offenbach, Germany. Previous research fields were evolutionary algorithms, biologically-inspired vision systems, information processing with spiking neurons and learning of sensory-motor maps. Since 2003 he is 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 Human-Machine Interaction project. He received M.S. 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 a car, 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 the scientific pursuit and technological innovation to create intelligent machines having associative interacting intelligence with humans in real world situations. He was the Director of the Japanese Society for Artificial Intelligence. He is a member of the IEEE, INNS, Society for Neuroscience, Robotic Society of Japan, Japan Society for Software and Technology and the Japanese Society for Artificial Intelligence.

Jun-ichi Imura was born in Gifu, Japan, in 1964. He received the M.S. degree in applied systems science, and the Ph.D. 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 IEEE, SICE, ISCIE, IEICE, and The Robotics Society of Japan.