Robust Ego Noise Suppression of a Robot

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

2010

Preprint:

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

Robust Ego Noise Suppression of a Robot

Gökhan Ince [1,3], Kazuhiro Nakadai [1,3], Tobias Rodemann [2],
Hiroshi Tsujino [1], and Jun-ichi Imura [3]

[1]Honda Res. Inst. Japan Co., Ltd. 8-1 Honcho, Wako-shi, Saitama 351-0188, Japan,
{gokhan.ince,nakadai,tsujino}@jp.honda-ri.com
[2]Honda Res. Inst. Europe GmbH, Carl-Legien Strasse 30, 63073 Offenbach, Germany,
tobias.rodemann@honda-ri.de
[3]Dept. of Mech. and Env. Informatics, Grad. School of Information Science and Eng.,
Tokyo Inst. of Tech. 2-12-1-W8-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan,
imura@mei.titech.ac.jp

Abstract. This paper describes an architecture that can enhance a robot with the capability of performing automatic speech recognition 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 template subtraction block and (3) a speech recognition block. In this work, we specifically investigate a missing feature theory based automatic speech recognition (MFT-ASR) approach in block (3), that makes use of spectrotemporal elements that are derived from (1) and (2) to measure the reliability of the audio features and to generate masks that filter unreliable speech features. We evaluate the proposed technique on a robot using word error rates. Furthermore, we present a detailed analysis of recognition accuracy to determine optimal parameters. Proposed MFT-ASR implementation attains significantly higher recognition performance compared to the performances of both single and multi-channel noise reduction methods.

Key words: Ego noise, noise reduction, robot audition, speech recognition, missing feature theory, microphone array

1 Introduction

Robots with microphones (such as NEC Papero [1]) are usually equipped with adaptive noise cancellation, beamforming and acoustic echo cancellation methods for robust speech recognition in noisy environments. However, the robot's own noise, so called ego noise, can also cause mis-recognition of spoken words during an interaction with a human, even if there are no other interfering sound sources in the environment. One special type of ego noise, which is observed while the robot is performing an action using its motors, is called *ego-motion noise*. This type of interference is more difficult to cope with compared to background noise or static fan-noise of the robot, because it is non-stationary and, to a certain extent, similar to the signals of interest. Therefore, conventional noise

reduction methods like spectral subtraction do not work well in practice. Several researchers tackled ego-motion noise problem by predicting and subtracting ego-motion noise using templates recorded in advance [2], [3]. Ince *et al.* [4] proposed to use a parameterized template subtraction which incorporates tunable parameters to deal with variations in the ego-motion noise. However, all those methods suffer from the distorting effects of *musical noise* [5] that comes along with nonlinear single-channel based noise reduction techniques and reduces the intelligibility and quality of the audio signal. Besides, this method has the following weakness: when applied together with a nonlinear stationary background noise prediction technique, e.g. Minima Controlled Recursive Averaging (MCRA) [6], in order to cope with the dynamically-changing environmental factors (e.g. stationary noise, reverberation), it creates a series of two consecutive nonlinear noise reduction operations. These operations produce even more musical noise, eventually causing damaged acoustic features and deteriorated recognition performance of automatic speech recognition (ASR).

It was shown that unreliable speech features degrade recognition performance severely [7]. Missing feature theory (MFT), which can be basically desribed as a filtering operation applied to the missing or damaged acoustic features, has already found useful applications like recognition of speech corrupted by music and several types of noise (refer to [7] for a comprehensive study) or simultaneous speech recognition of several speakers in the field of robot audition [8], [9], where they based their models of mask generation on the disturbing effect of leakage noise over speech caused by an imperfect source separation. In this work, we incorporate MFT to solve the ego-motion noise problem of a robot. To estimate the reliability of the features of speech, which is subject to residuals of motor noise after template subtraction and to improve the performance of ASR, we propose to use MFT with a model that is based on the ego-motion noise estimations. To generate suitable masks, we propose to integrate also a multi-channel framework that consists of sound source localization (SSL), sound source separation (SSS), and speech enhancement (SE), of which the first two steps make use of the directivity properties of motor noises to cancel them and thus provide additional information about the reliability assessment. In this respect, the main contribution of our work will be the incorporation of an original missing feaute mask (MFM) generation method based on the signals available from 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. Firstly, a binary mask, that uses either 0 or 1 to estimate the reliability of each acoustic feature, is suggested. We, later, enhance the proposed method further by using a soft mask represented as continuous values between 0 and 1 that yields more detailed information about the reliability. We demonstrate that the proposed methods achieve a high noise elimination performance and thus ASR accuracy.

The rest of the paper is organized as follows: Section 2 describes the proposed system and briefly summarizes the preprocessing-stages, namely SSL, SSS, SE and template subtraction. Section 3 investigates speech recognition integration

and computation of the missing feature masks in detail. Conducted experiments and consecutive results are presented in Section 4. The last section gives a conclusion and future work.

2 System Overview

As sensors we use 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 in the environment. The estimated locations of the sources are used by a linear separation algorithm called Geometric Source Separation (GSS) [8]. It can be considered as a hybrid algorithm that exerts Blind Source Separation (BSS) [10] and beamforming. The next stage after SSS is a speech enhancement step called multichannel Post Filtering (PF). 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. These three main modules constitute the **multi-channel noise reduction** block [11], whereas the second block performs **template subtraction** [4]. Altogether, both branches are responsible for the *audio features* for speech recognition and *spectrograms* to be processed further 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 in order to extract the most suitable features. This part will be discussed in Section 3 in detail.

2.1 Multi-channel Noise Reduction System [11]

In order to estimate the Directions of Arrival (DoA) of each sound source, we use a popular adaptive beamforming algorithm called MUltiple Signal Classification (MUSIC) [12]. It detects the 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 finding the peaks occurring in the spatial spectrum. A consequent source tracker system performs a temporal integration in a given time window.

Geometric Source Separation [10], later on extended to be an adaptive algorithm that can process the input data incrementally [13], makes use of the locations of the sources explicitly. To estimate the separation matrix properly, GSS introduces cost functions that must be minimized in an iterative way (see [13] for details). Moreover, we use adaptive step-size control that provides fast convergence of the separation matrix [14].

After the separation process, a multi-channel post filtering operation proposed by Valin [13] is applied, which can cope with nonstationary interferences as well as stationary noise. This module treats the transient components in the spectrum as if they are caused by the leakage energies that may occasionally

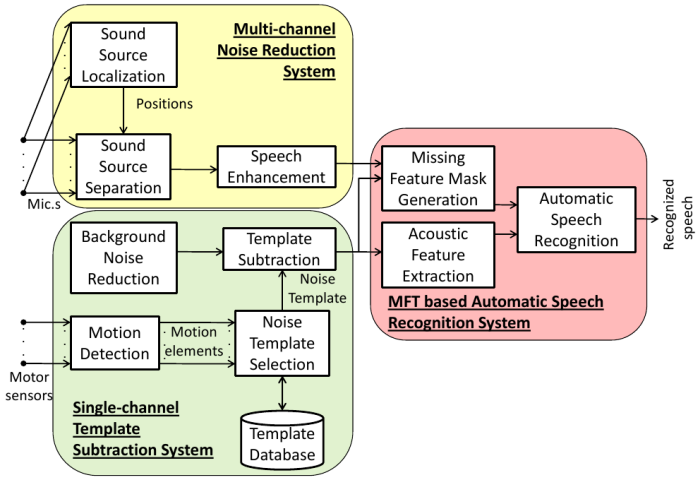

Fig. 1. Proposed noise cancellation system

arise due to poor separation performance. For this purpose, noise variances of both stationary noise and source leakage are predicted. Whereas the former one is computed using the MCRA [5] method, to estimate the latter the algorithm proposed in [13] is used.

2.2 Single-channel Template Subtraction System [4]

During the motion of the robot, current position (θ) information from each joint is gathered regularly in the template generation (database creation) phase. Using the difference between consecutive motor position outputs, velocity ($\dot{\theta}$) values are calculated, too. Considering that J joints are active, joint position vectors with the size of $2J$ are generated. The resulting vector has the form of $F{=}[\theta_1, \dot{\theta}_1, \theta_2, \dot{\theta}_2 \ldots, \theta_J, \dot{\theta}_J]$. At the same time, motor noise is recorded and the spectrum of the motor noise is calculated in parallel with motion element acquisition. Both joint position vectors and spectra are continuously labeled with time tags so that they can be synchronized. Finally, a large noise template database that consists of short noise templates for desired joint configurations is created. In the prediction phase a nearest neighbor search in the database is conducted for the best matching template of motor noise for the current time instance (frame at that moment) using the joint-status vectors. The templates are used as weights inside the spectral subtraction routine.

3 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. As

stated in [7], Missing Feature Theory is a very promising approach that basically applies a mask to decrease the contribution of unreliable parts of distorted speech. By keeping the reliable parameters that are essential for speech recognition, a substantial increase in recognition accuracy is achieved [8], [9]. In this section, we will discuss the basic steps of such an ASR system and how this approach can be adapted to fit to the ego-motion noise problem by presenting a robust mask design method for estimating reliability of speech based on the current motor noise.

3.1 Acoustic Feature Extraction

Acoustic features are extracted from the refined spectrum, which is the final product of the noise reduction stage (See Fig. 1). Because we do not want to have the distortions spreading to all coefficients of the cepstrum, we avoided the usage of Mel-Frequency Cepstral Coefficients (MFCC) in contrast to conventional ASR systems. Instead, we used the Mel-Scale Log Spectrum (MSLS), whose detailed calculation method can be found in [15]. Moreover, linear regression of each spectral coefficient is represented as a *delta* feature and it is used to enhance the quality of acoustic features. A consequent stage of spectral mean normalization improves noise robustness of MSLS features by subtracting the average of the features in the last 5 sec. from the current features.

3.2 MFM Generation

The reliability of features is computed for each frame and for each mel-frequency band. If continuous values between 0 and 1 constitute the mask, it is called a *soft mask*. On the other hand, a *hard mask* contains only discrete values, either 0 or 1. In this paper, we used both methods to assess their performance for this particular type of ego-noise problem. We start by explaining some underlying findings about the ego-motion noise suppression capabilities of the preprocessing stages of our proposed system in the next two paragraphs as a motivation. Then, we show how to derive the masks later in detail.

GSS lacks the ability to catch motor noise originating from the same direction of the speaker and suppress it, because the noise is considered as part of the speech. Moreover, when the position of the noise source is not detected precisely, GSS cannot separate the sound in the spatial domain. As a consequence, motor noise can be spread to the separated sound sources in small portions. However, multi-channel noise suppression systems work very well for weaker motion noises like arm or leg motions compared to head motion noise, as we found out in our experiments [11]. Additionally, it is optimally designed for "simultaneous multiple speakers" scenarios with background noise and demonstrates a very good performance when no motor noise is present.

On the other hand, template subtraction does not make any assumption about the directivity or diffuseness of the sound source and can match a prerecorded template of the motor noise at any moment. The drawback of this

approach is, however, due to the non-stationarity, the characteristics of predicted and actual noise can differ to a certain extent.

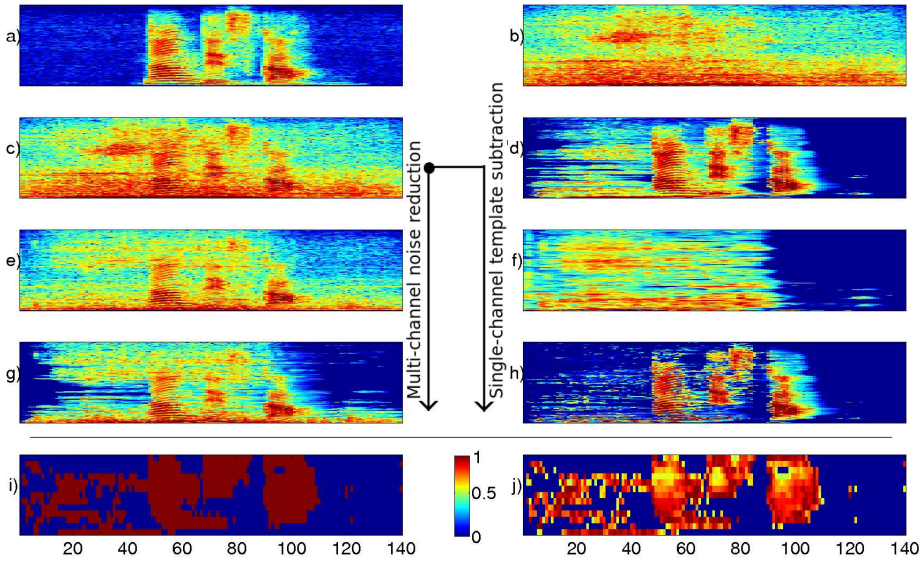

Fig. 2. Spectrogram of (a) clean speech, (b) motor noise + background noise, (c) noisy speech (**a**+**b**), (d) background noise reduction (MCRA) applied to **c**, (e) GSS applied to **c**, (f) extracted template for template subtraction, (g) PF applied to **e**, (h)template subtraction applied to **d** using **f**, (i) hard mask generated using **g** and **h**, (j) hard mask generated using **g** and **h**. In (a)-(h), y-axis represents frequency bins between 0 and 8kHz, in (i)-(j) 13 static mel-features are represented in y-axis. x-axis represents in all panels the index of frames.

As we have stated, the strengths and weaknesses of both approaches are distinct and thus can 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. Computation of the masks is performed for each frame, k, and for each frequency band, f. First, a continuous mask is calculated like following:

$$m(f,k)=\frac{|\hat{S}_m(f,k)-\hat{S}_s(f,k)|}{\hat{S}_m(f,k)+\hat{S}_s(f,k)}, \tag{1}$$

where $\hat{S}_m(f,k)$ and $\hat{S}_s(f,k)$ are the estimated energy of the refined speech signals, which were subject to multi-channel noise reduction and resp. single-channel template subtraction. Both signals are computed using a mel-scale filterbank. The numerator term represents the deviation of the two outputs, which is a measure of the uncertainty or unreliability. The denominator term, however,

is a scaling constant and is given by the average of the two estimated signals. (To simplify the equation, we remove the scalar value in the denominator, so that $m(f,k)$ can take on values between 0 and 1.) Depending on the type of the mask (hard or soft) used in the MFT-ASR, Eq.(2) or Eq.(3) is selected.

1. For hard (binary) mask:

$$M(f,k)=\begin{cases}1, & \text{if } m(f,k)<T\\ 0, & \text{if } m(f,k)\geq T\end{cases}. \tag{2}$$

2. For soft mask [9]:

$$M(f,k)=\begin{cases}\dfrac{1}{1+\exp(-\sigma(m(f,k)-T))}, & \text{if } m(f,k)<T\\ 0, & \text{if } m(f,k)\geq T\end{cases}, \tag{3}$$

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

Fig. 2 gives a general overview about the effect of each processing stage until the masks are generated. In Fig. 2c), we see a tightly overlapped speech (Fig. 2a)) and motor noise (Fig. 2b)) mixture with an SNR of -5dB. GSS+PF in Fig. 2g) reduces only a minor part of the motor noise while sustaining the speech. On the other hand, template subtraction (Fig. 2h)) reduces the motor noise aggresively while damaging some parts of the speech, where some features of the speech get distorted. The hard mask (Fig. 2i)) gives us a filter eliminating unreliable and still noisy parts of the speech (T=0.5). The soft mask (Fig. 2j)), in addition, provides more detailed information about the reliability degree of each feature so that the noise-free features are weighted more than the noise-containing parts in the MFT-ASR ($\{T,\sigma\}$=$\{0.5,5\}$). Furthermore, we observe that weights in the first 50 frames contaminated with noise were given either zero or low weights in the mask. Note that speech features are located between the [50 110]-th frames.

3.3 MFT-ASR

Missing Feature Theory Based Automatic Speech Recognition (MFT-ASR) is a Hidden Markov Model based speech recognition technique [7]. Suppose $M(i)$ is the MFM vector that is generated as in Sec. 3.2 for the i-th acoustic feature, the output probability can be given as follows:

$$b_j(x)=\sum_{l=1}^{L}P(l|S_j)\exp\left\{\sum_{i=1}^{I}M(i)\log f(x(i)|l,S_j)\right\}, \tag{4}$$

where $b_j(x)$ is the output probability of j-th state, $x(i)$ is denoted as an acoustic feature vector, I represents 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. Please note that when all mask values are set to 1, Eq.(4) becomes the same as the output probability calculation of a conventional ASR.

4 Results

In this section we present comparative results for pre-processing based ASR, hard and soft mask based ASR, and the influence of selected parameters for template subtraction and MFT-ASR on the performance. To evaluate the performance of the proposed techniques, we use Honda's humanoid robot ASIMO. The robot is equipped with an 8-ch microphone array on top of its head. Of the robots many degrees of freedom, we use only 2 motors for head motion, and 4 motors for the motion of each arm with altogether 10 degrees of freedom. We recorded random motions performed by the given set of limbs by storing a training database of 30 minutes and a test database 10 minutes long. Because the noise recordings are comparatively longer than the utterances used in the isolated word recognition, we selected those segments, in which all joints contribute to the noise. The noise signal consisting of ego noise (incl. ego-motion noise) and environmental background noise is mixed with clean speech utterances used in a typical human-robot interaction dialog. This Japanese word dataset includes 236 words for 4 female and 4 male speakers. Acoustic models are trained with Japanese Newspaper Article Sentences (JNAS) corpus, 60-hour of speech data spoken by 306 male and female speakers, hence the speech recognition is a word-open test. We used 13 static MSLS, 13 delta MSLS and 1 delta power. Speech recognition results are given as average WER of instances from the test set. In 'isolated word recognition' tasks, WER is an evaluation criteria alternative to Word Correct Rate (WCR), such that $WER = 100\% - WCR$ holds. The position of the speaker is kept fixed at 0° throughout the experiments. The recording environment is a room with the dimensions of 4.0 m×7.0 m×3.0 m with a reverberation time (RT_{20}) of 0.2s. Although the position of the original sound source was given in advance to avoid the 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.

Fig. 3a) illustrates the ASR accuracies for all methods under consideration. The results are evaluated using an acoustic model trained with MCRA-applied speech data, except GSS+PF method for which we used a matched acoustic model for that condition. We evaluated MFMs for three heuristically selected threshold parameters T=$\{0.25, 0.5, 0.75\}$. In the preliminary tests we found out that the feature set that is derived at the output of template subtraction achieves higher accuracy in a range of 10 to 20% in WER compared to the features after multi-channel noise reduction. So, we concluded that the former feature type is more suitable to be used in an MFT-ASR. Single-channel results are used as a baseline. MFM-ASR outperforms both mere single (TS) & multi-channel (GSS+PF) noise reduction methods. In the case of T<0.5, indicating a mask model based on a reliability estimation exerting sharp differentiation of both evidences of noise reduced spectra was unable to improve the ASR, because essential features belonging to the speech are thrown away, thus WERs deteriorate. On the contrary, higher thresholds improved the outcomes significantly.

In the second part of our experiments, we compared the results of hard masking with optimal threshold (T=0.75) obtained in the first part of the ex-

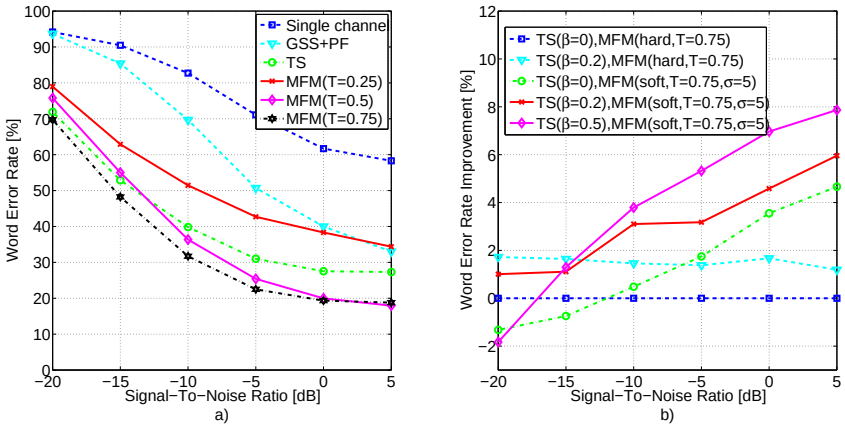

Fig. 3. Speech recognition performance for a) different processing stages b) soft mask - hard mask comparison for given parameters.

periments, to the results of soft masking for a parameter set of σ=$\{5, 10, 50\}$. All three cases with the given parameters yielded more or less the same WER improvements, however, outside this range the results become very sensible to σ and worsen eventually. Therefore, we will only present the results for σ=5. Besides, we inspected the effect of decreasing the aggresiveness level of the template subtraction, by leaving an artificial floor on the bottom of the spectra. So far, the parameter called *spectral floor* (β, where $0 \leq \beta \leq 1$) [4] was set to zero. We assess the results for β=$\{0, 0.2, 0.5\}$ in the framework of soft-hard mask comparison in Fig. 3b) by giving the WER improvement relative to the hard mask results obtained for β=0 and T=0.75. By increasing β, we observed that the WERs improve considerably. That means that a tradeoff between "noise reduction level" and "signal distortion" contributed to the mask quality substantially. Furthermore, soft masks reduce the WERs even further by up to 8% compared to hard masks. This reduction is attained due to the improved probabilistic representation of the reliability of each feature. Optimal results are obtained when we use a soft mask with the following parameter set: $\{T, \sigma, \beta\}$=$\{0.75, 5, 0.5\}$.

5 Summary and Outlook

In this paper we presented a method for eliminating ego-motion noise from speech signals. The system we proposed to utilize (1) a multi-channel noise reduction stage, (2) a template subtraction scheme, and finally (3) a masking stage to improve speech recognition accuracy. We used an MFM model, which is based on the similarity measurements of ego-motion noise estimations gathered from (1) and (2). We validated the applicability of our approach by evaluating its performance for different settings for both hard and soft MFM. Our method demonstrated significant WER improvement for hard masking (45% relative to single-channel recognition) and soft masking (up to 53%).

In future work, we plan to find an optimized parameter set for template subtraction and especially for MFM-ASR block in a wider range. The next step is an evaluation in real time and a real situation, which involves speech recognition of several speakers simultaneously while the robot is performing some motion.

References

1. Sato, M., Sugiyama, A., and Ohnaka, S.: An adaptive noise canceller with low signal-distortion based on variable stepsize subfilters for human-robot communication. IEICE Trans. Fundamentals **E88-A** 8 (2004) 2055-2061
2. Nishimura, Y., Nakano, M., Nakadai, K., Tsujino, H., and Ishizuka, M.: Speech Recognition for a Robot under its Motor Noises by Selective Application of Missing Feature Theory and MLLR. ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition (2006)
3. Ito, A., Kanayama, T., Suzuki, M., Makino, S.: Internal Noise Suppression for Speech Recognition by Small Robots, Interspeech (2005) 2685-2688
4. Ince, G., Nakadai, K., Rodemann, T., Hasegawa, Y., Tsujino, H., Imura, J.: Ego Noise Suppression of a Robot Using Template Subtraction. Proc. of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS) (2009) 199-204
5. Cohen, I.: Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement. IEEE Signal Processing Letters **9** 1 (2002)
6. Boll, S.: Suppression of Acoustic Noise in Speech Using Spectral Subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing, **ASSP-27** 2 (1979)
7. Raj, B., Stern,R. M.: Missing-feature approaches in speech recognition. IEEE Signal Processing Magazine **22** (2005) 101-116
8. Yamamoto, S., Nakadai, K., Nakano, M., Tsujino, H., Valin, J. M., Komatani, K., Ogata, T., Okuno, H. G.: Real-time robot audition system that recognizes simultaneous speech in the real world. Proc. of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS) (2006)
9. Takahashi, T., Yamamoto, S., Nakadai, K., Komatani, K., Ogata, T., Okuno, H. G.: Soft Missing-Feature Mask Generation for Simultaneous Speech Recognition System in Robots. Proceedings of International Conference on Spoken Language Processing (Interspeech) (2008) 992-997
10. Parra, L. C., Alvino, C. V.: Geometric Source Separation: Merging Convolutive Source Separation with Geometric Beamforming. IEEE Trans. Speech Audio Process. **10** 6 (2002) 352-362
11. Ince, G., Nakadai, K., Rodemann, T., Hasegawa, Y., Tsujino, H., Imura, J.: A Hybrid Framework for Ego Noise Cancellation of a Robot. Proc. of the IEEE/RSJ International Conference on Robotics and Automation (ICRA) (2010) to appear
12. Schmidt, R., Multiple emitter location and signal parameter estimation. IEEE Trans. on Antennas and Propagation **34** 3 (1986) 276-280
13. Valin, J.-M., Rouat, J., Michaud, F.: Enhanced Robot Audition Based on Microphone Array Source Separation with Post-Filter. Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2004) 2123-2128
14. Nakajima, H.,Nakadai, K., Hasegawa, Y., Tsujino, H.: Adaptive step-size parameter control for real-world blind source separation. Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2008) 149-152
15. Nishimura, Y., Shinozaki, T., Iwano, K., Furui, S.: Noise-robust speech recognition using multi-band spectral features. Proc. of 148th Acoustical Society of America Meetings **1aSC7** (2004)