Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

A Hybrid Framework for Ego Noise Cancellation of a Robot

Gökhan Ince, Kazuhiro Nakadai, Tobias Rodemann, Yuji Hasegawa, 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]

A Hybrid Framework for Ego Noise Cancellation of a Robot

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

***Abstract*— Noise generated due to the motion of a robot is not desired, because it deteriorates the quality and intelligibility of the sounds recorded by robot-embedded microphones. It must be reduced or cancelled to achieve automatic speech recognition with a high performance. In this work, we divide ego-motion noise problem into three subdomains of arm, leg and head motion noise, depending on their complexity and intensity levels. We investigate methods that make use of single-channel and multi-channel processing in order to suppress ego noise separately. For this purpose, a framework consisting of a microphone-array-based geometric source separation, a consequent post filtering process and a parallel module for template subtraction is used. Furthermore, a control mechanism is proposed, which is based on signal-to-noise ratio and instantaneously detected motions, to switch to the most suitable method to deal with the current type of noise. We evaluate the proposed techniques on a humanoid robot using automatic speech recognition (ASR). The preliminary results of isolated word recognition show the effectiveness of our methods by increasing the word correct rates up to 50% compared to the single channel recognition in arm and leg motion noises and up to 25% in very strong head motion noises.**

I. Introduction

In daily environments, where robots are intended to be employed in the near future, a lot of noise sources exist. Therefore, a robot audition system must be able to cope with all kinds of noises including the robot's own noises, i.e. ego noises, during an interaction with a human. 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 noise is rather ignored [1] or circumvented by using close-talk microphones [2] in the robotics literature, however with increasing popularity and growing demand on home/service robots, it will apparently become an important problem.

Nakadai *et al.* [3] proposed a noise cancellation method with 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 the spectral subbands that are noisy and not noisy, and to ignore the ones where the noise is dominant. Besides, some single-channel based approaches are introduced to deal with ego-motion noise like the following studies: Nishimura *et al.* [4] estimated the ego-noise using robot's gestures and motions. With the help of the motion command, the pre-recorded correct noise template matching to the recent motion was selected from the template database and subtracted. Ito *et al.* [5] developed a new approach of frame-by-frame based prediction with a neural network to cope with unstable walking noise. The trained network had to predict the noise spectrum from angular velocities of the joints of the robot. In another work, analysis results of ego-motion noise [6] showed clearly that it has a highly non-stationary nature. Therefore, Ince *et al.* [6] proposed to use template subtraction which incorporates tunable parameters to cope with noise template representations that do not match to the instantaneous noise due to the deviations in the noise spectra. However, all those methods suffered from the *musical noise* [7], which can be described as smaller attenuations of the frequencies compared to relatively larger attenuations of their neighboring frequencies caused by non-linear mapping of the negative or small-valued spectral estimates. This distorting effect comes along with nonlinear single-channel based noise reduction techniques and reduces the intelligibility and quality of the audio signal. If we consider also that in order to cope with the dynamically-changing environmental factors such as background noises and unknown source positions, we apply a nonlinear stationary background noise reduction technique, e.g. Minima Controlled Recursive Averaging (MCRA) [8] prior to ego-motion noise reduction. Two consecutive nonlinear noise reduction operations produce even more musical noise, eventually causing deteriorated recognition performances of automatic speech recognition (ASR).

In this work, we propose the use of a framework that consists of a microphone array, sound source localization (SSL), sound source separation (SSS), speech enhancement (SE) and template subtraction to cancel motor noises. Furthermore, ASR is integrated to the framework to evaluate the results of each processing stage qualitatively. Because ego-motion noise is created in the near field of the microphone array, we assume that it is not only a directional, but also a diffuse type of noise. To tackle the directional portion of the ego noise, we utilize the SSS. We also apply spectral enhancement techniques, because they are the most suitable way to deal with the diffuse portion of the noise. To our knowledge, ego-motion noises are never tackled by using a multi-channel sound source separation and post filtering technique before, which makes this study also a proof of concept for multi-channel ego noise reduction. Moreover, we disaggregated the whole body motion ego-noise problem

Gökhan Ince, Kazuhiro Nakadai, Yuji Hasegawa and Hiroshi Tsujino are with Honda Research Institute Japan Co., Ltd. 8-1 Honcho, Wako-shi, Saitama 351-0188, Japan {gokhan.ince, nakadai, yuji.hasegawa, tsujino}@jp.honda-ri.com

Tobias Rodemann is with Honda Research Institute Europe GmbH, Carl-Legien Strasse 30, 63073 Offenbach, Germany tobias.rodemann@honda-ri.de

Gökhan Ince, Kazuhiro Nakadai, Jun-ichi Imura are with Dept. 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 imura@mei.titech.ac.jp

mainly into three categories that can be analyzed separately from each other and investigate the performance of multi-channel approach for each of them. The main contribution of our work will be incorporation of the SSS stage for a smooth speaker/ego-noise separation and utilization of the SE stage for ego-motion noise suppression. We also enhance the proposed system further by incorporating template subtraction method into the hybrid framework to compansate the poor performance of multi-channel approach especially with the head motion noise (See Fig. 1). We demonstrate that the proposed methods achieve a high noise elimination performance and thus improve speech recognition accuracy.

The rest of the paper is organized as follows: Section II describes an overview of the system. Section III presents the main building blocks of the proposed framework that is composed of SSL, SSS, SE and template subtraction stages in detail. Section IV shows the conducted experiments and consecutive results. The last section gives a conclusion and future work.

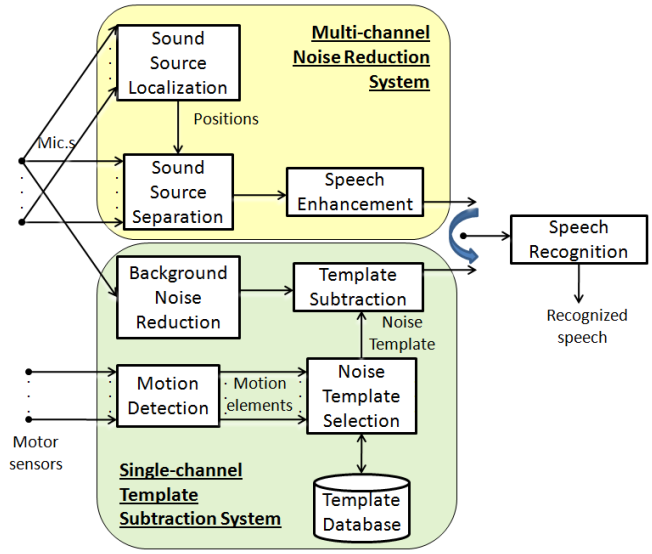

Fig. 1. Proposed hybrid noise cancellation system. The blue arrow implies a switch between two separate systems that operate simultaneously.

II. System Overview

We propose to use an array that consists of multiple omnidirectional microphones and that is mounted on the robot for this approach. The first building block of our processing chain is composed of the elements performing SSL that extracts the location information of the most dominant sources in the environment. Basing on the selection of the value assigned to the threshold parameter embedded in this module (see Sec. III-A), the number of detected sources can vary in time and space. The estimated locations of the sources are used by a linear separation algorithm called Geometric Source Separation (GSS) [9]. It is a hybrid algorithm that exerts Blind Source Separation (BSS) [10] and beamforming. This method has three important advantages for the ego-noise cancellation problem.

1) The introduction of geometric constraints concept that involves calculation of current transfer function based on the known locations of the microphones and positions of the sound sources obtained from SSL relaxes the limitations of BSS such as permutation and scaling problems. Therefore it can run in real-time.
2) Sound separation of moving sources is possible. This is especially important if we consider that the part of the robot where the microphones are mounted (e.g. head) can move as well. Relative to a moving microphone array, even stationary sound sources are regarded as moving objects.
3) Generally, an embodied robot has loud ego noises such as stationary operational noise of hardware and fan noise, which are also located close to each other. Assuming we know the positions of these high noise emission sources, we can specify their direction, because our GSS module has a function of suppressing stationary ego noise as a fixed noise source.

The next stage after SSS is a speech enhancement step called multichannel Post Filtering (PF). This block attenuates stationary noises, e.g. background noise, and non-stationary noises that arise because of the leakage energy between the output channels of the previous separation stage for each individual sound source. We also inspected single-channel template subtraction module's performace as an alternative to the multi-channel approach. The overall architecture of the proposed noise reduction system is shown in Fig. 1.

As a final operation, the appropriate features are extracted from the output of either PF or template subtraction operation, which represent the inputs of the ASR module.

III. System Architecture

For our multi-channel approach, we will use the following signal model for M sources and N ($\geq M$) microphones throughout the text: $\mathbf{X}(\omega) = [X_1(\omega), X_2(\omega), \cdots, X_N(\omega)]$ with $X_n(\omega)$ being the spectrum of the signal captured by the n-th microphone. ω denotes the angular frequency.

The following subsections explain processing blocks of SSL, SSS, PF and template subtraction in detail.

A. Sound Source Localization

In order to estimate the directions of arrival (DoA) of the sound sources, we will use one of the most popular adaptive beamforming algorithms called MUltiple SIgnal Classification (MUSIC) [11]. It detects the DoA by performing an eigenvalue decomposition on the correlation matrix of the noisy signal such as following:

$$\mathbf{R}_{xx}(\omega, \phi) = \mathbf{X}(\omega)\mathbf{X}^*(\omega), \tag{1}$$

where $()^*$ represents complex conjugate transpose operator and ϕ denotes the orientation of the robots head. Eigen decomposition of $\mathbf{R}_{xx}(\omega, \phi)$ leads to

$$\mathbf{R}_{xx}(\omega, \phi) = \mathbf{Q}(\omega, \phi)\Lambda\mathbf{Q}^{-}1(\omega, \phi), \tag{2}$$

where Λ is the matrix, whose diagonal elements are the corresponding eigenvalues, i.e. $\Lambda_{ii} = \lambda_i$ and $\mathbf{Q}$ is the square matrix, whose i-th column is the eigenvector $\mathbf{q}_i$. Moreover,

we assume that the λ_i and $\mathbf{q}_i$ belong to the sound sources of interest for $1 \leq i \leq M$ and to the undesired noise sources for $M+1 \leq i \leq N$.

Prior to localization, steering vectors of the microphone array, $\mathbf{G}(\omega, \psi)$, are determined, which are measured as impulse responses for a certain orientation of ψ.

$$\mathbf{P}(\omega, \psi) = \frac{|\mathbf{G}^*(\omega, \psi)\mathbf{G}(\omega, \psi)|}{\sum_{n=M+1}^{N} |\mathbf{G}^*(\omega, \psi)\mathbf{q}_n|}. \quad (3)$$

The peaks occurring in the spatial spectrum yield the source locations. Moreover, a consequent source tracker system, which actually performs a temporal integration of the source directions in a given time window, runs to ensure the reliability of the location estimations. The decision on the source locations is made by comparing the power of the peaks of $\mathbf{P}(\omega, \psi)$ to a threshold value T and if the power of the source is less than the threshold, the source is eliminated. Currently, we set the threshold manually.

B. Sound Source Separation

We present here Geometric Source Separation which is an adaptive algorithm that can process the input data incrementally and makes use of the locations of the sources explicitly. It requires lower computational cost compared to ICA-based BSS algorithms.

Suppose $\mathbf{W}(\omega)$ is the separation matrix, separated sources $\mathbf{Y}(\omega)$ can be found such as below:

$$\mathbf{Y}(\omega) = \mathbf{W}(\omega)\mathbf{X}(\omega). \quad (4)$$

To estimate $\mathbf{W}(\omega)$ properly, GSS introduces cost functions that must be minimized in an iterative way (Refer to [12] for details). Moreover, we use adaptive step-size control that provides fast convergence of the separation matrix [13]. Besides, our GSS implementation also exploits a method called Optima Controlled Recursive Averaging [14], which controls window size adaptively causing a smoother convergence and thus better separation results [15].

C. Speech Enhancement

After the separation process, a multi-channel post filtering operation is applied so that the sounds can be enhanced further. This module is based on the optimal estimator proposed by Ephraim and Malah [16]. Since their method takes temporal and spectral continuities into consideration, it generates less distortion compared to the conventional spectral subtraction based noise reduction methods. By extending their idea further, a multichannel post filter is proposed by Cohen [17], 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 arise due to poor separation performance .

The main aim of post filtering is to find the weighting coefficients $G_m(\omega)$ and estimate the clean audio signal that is represented by $\hat{S}_m(\omega)$ by attenuating $Y_m(\omega)$ as in Eq. (5).

$$\hat{S}_m(\omega) = G_m(\omega)Y_m(\omega). \quad (5)$$

For this purpose, noise variances of both stationary noise $\lambda_m^{stat}(\omega, n)$ and source leakage $\lambda_m^{leak}(\omega, n)$ must be predicted. Whereas the former one is computed using the MCRA [8] method, to estimate the latter $\lambda_m^{leak}(\omega, n)$ the formulations proposed in [12] are used. The noise suppression rule further involves speech presence probability calculations such as given in [17] and is based on minimum mean-square error estimation of the spectral amplitude [16]. According to the outcomes of our experiments, we conclude heuristically that an eventual additive white noise step applied after post filtering improves the speech recognition results by generating an artificial spectral floor in the background of a speech signal and blurring the musical noise distortions.

D. Template Subtraction [6]

This method requires sensors attached to each motor (joint) to measure its angular positions individually. This noise reduction method works like the following: During the motion of the robot, actual position (θ) information regarding each motor is gathered regularly in the template generation (database creation) phase. Using the difference between consecutive sensor outputs, velocity ($\dot{\theta}$) values are calculated, too. Considering that N joints are active, feature vectors with the size of $2N$ are generated. The resulting feature vector has the form of $F = [\theta_1, \dot{\theta}_1, \theta_2, \dot{\theta}_2 \ldots, \theta_N, \dot{\theta}_N]$. At the same time, motor noise is recorded and spectrum of the motor noise is calculated by the sound processing branch running in parallel with motion element acquisition. Both feature vectors and spectra are continuously labeled with time tags so that templates are generated when their time tags match. Finally, a large noise template database consisting of short noise templates for many 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 using the feature (joint-status) vectors. The coefficients are calculated from the selected templatess for the weighting operation in a similar fashion like in Eq. (5).

IV. Results

In order to evaluate the performance of the proposed multi-channel approach, we used ASIMO. As depicted in Fig. 2, the robot is equipped with an 8-ch microphone array, 2 motors for head motion, 4 motors for the motion of each arm, 5 motors to move each leg.

It is clear that using the above-mentioned microphone array configuration the neck motors are the closest sound sources, thus the most problematic ones, because the intensity of a sound wave depends on how far it is from a source with the basic formula:

$$SoundIntensity = SoundPower/(4\pi R^2), \quad (6)$$

where R denotes the distance. Therefore, we decided to handle the noise problem in different domains, each one covering a set of joints required for a certain type of an interaction with the robot's environment. We recorded random motions performed by a given set of limbs, which

can be classified mainly into 3 distinct categories following the order of increasing noise intensity: *arm motion*, *leg motion* and *head motion*.

Fig. 2. Experiments are conducted on ASIMO whose legs, arms and head can move. Motion noise is recorded by an 8-ch microphone array with a circular layout embedded on ASIMO's head.

Because the noise recordings are comparatively longer than the utterances used in the isolated word recognition, we selected especially those segments, in which all contributing joints of corresponding category were active, thus the noisiest parts of the recordings. The noise signal consisting of ego noise (incl. ego-motion noise) and environmental background noise is mixed with clean speech utterances used in a daily human-robot interaction dialog. This Japanese word dataset includes 236 words per 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. Furthermore, multicondition training of an acoustic model is performed for each processing technique to be able to compare the results of each processing stage in a better way. Speech recognition accuracy on clean audio files is around 97%. Speech recognition results are given as average word error rates (WER) of five arbitrarily selected noise instances from corresponding noise categories. The position of the speaker is kept fixed at 0° throughout the experiments. Besides, recording environment was a room with the dimensions of $4.0\,\mathrm{m}\times 7.0\,\mathrm{m}\times 3.0\,\mathrm{m}$ with a reverberation time (RT_{20}) of 0.2s. The implementation runs on HARK, which is an open-sourced software for robot audition [18].

A. Speech Recognition with Arm Motion Noise

While moving arms (whole-arm motion pointing behavior), the microphone array and the head are kept stationary. Henceforth, we are able to fix the direction of the ego-noise originating from the backpack of ASIMO (-180°). Note that giving a fixed ego-noise direction does not pose any hard constraint on robot audition scenario or application, because the robot is already equipped with sensors that transmit the positions of the joints. Depending on the posture of the body, we exactly know where the ego-noise is emitted and change the direction automatically.

The results are presented for five different conditions:

- Single channel recognition,
- GSS (implied as SSS) performed with a high threshold $T = 25dB$ (See Sec. III-A for the usage of T),
- GSS and Post Filter (implied as SE) with a low threshold $T = 23dB$,
- GSS and Post Filter with a high threshold $T = 25dB$,
- GSS and Post Filter with known source location.

Note that the threshold values are determined heuristically to ensure the accuracy of the detected source locations.

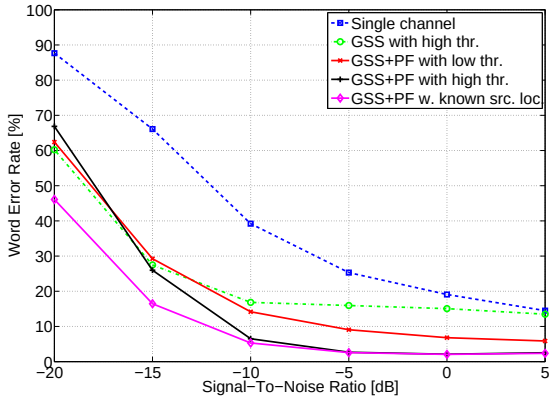

Fig. 3. Recognition performance of speech with arm motion noise.

Speech recognition accuracy results are shown in Fig. 3. Single-channel results are used as a baseline. As expected, the GSS+PF system achieved up to 40% improvement compared to the single-microphone based recognition and outperformed GSS by increasing the ASR rates by an additional 10%. This result proves that the arm-motion noise can be treated as a directional & diffuse non-stationary noise source that can be handled by GSS & PF stages. We also included GSS+PF, which makes use of the locations obtained from SSL with a low threshold, in order to show the importance of the threshold selection. If an inappropriately low threshold is selected, additional non-existing *ghost* sources are detected, which at the end deteriorates the performance of GSS and PF. On the other hand, GSS+PF with high threshold causes missing sources at low signal-to-noise ratio (SNR) cases that diminish the performance in another way. For an additional test bench, we also introduce "GSS+PF with known source location" results, where we assume that the location of the sound source is estimated precisely. Though it may seem that it achieves only a small improvement on the ASR accuracy, the result is significant, because it demonstrates the upper performance limit of our proposed method just in case we solve the SSL problem.

B. Speech Recognition with Leg Motion Noise

The legs are used for performing stamping behavior and short distance walking. Again, the same conditions as in the previous experiment are provided. The recognition results curves in Fig. 4 show very similar patterns as in Fig. 3. This time, we observe severely deteriorated outcomes for the GSS+PF method provided by an SSL that runs with a low

threshold. Because legs' noise level is considerably higher and even more complex than arms' noise, the localization system fails with an improper setting, thus yielding incorrect position information to the next processing stages. However, for an optimally tuned threshold value, drastically high suppression rates can be achieved even for high SNR's. The post filter contributes to a 30-50% reduction in the WERs.

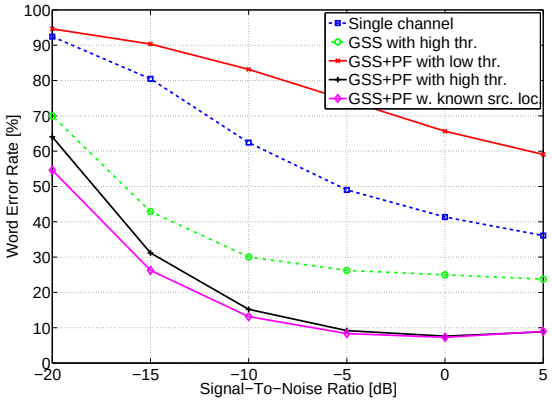

Fig. 4. Recognition performance of speech with leg motion noise.

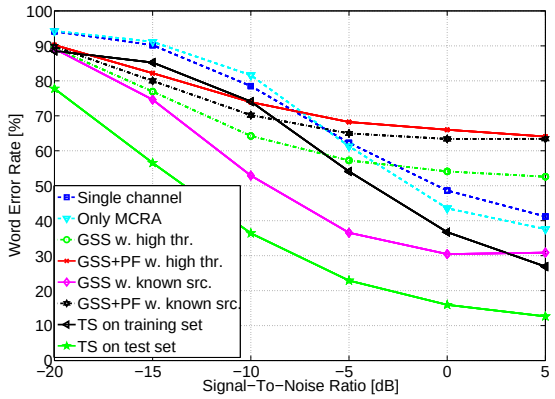

Fig. 5. Recognition performance of speech with head motion noise.

C. Speech Recognition with Head Motion Noise

Microphones' current placement provides the fact that whenever the head moves, the microphone array rotates as well. Another consequence of the head motion is of course, the relative motion of sound sources and ego-noise with respect to the microphones. Since in this work we only applied isolated word recognition, the effect of the moving sound sources on the separation and speech enhancement performance is rather mild. Nevertheless, to inspect the capabilities of our proposed noise reduction system based on the SSS and keep the results coherent with future research extensions of this work, we did not fix the ego-noise direction of the robot. In this experiment, SSL system predicted it automatically.

The head motor noise is extremely loud due to its close proximity. Our partial directional & diffuse noise assumption is violated, because a strong noise source in the very near field of the microphone array has highly complicated propagation pattern. As a consequence, the separation quality gets worsened and the noise model used in the post filtering stage also does not hold any more (e.g. the transient components in the separated signal spectrum are due to leakage energies, etc.). Hence, after validating the performance of the proposed multi-channel approach, we want to compare the results with those of single-channel template subtraction technique. This method does not model the noise depending on its nature, but rather uses instantaneous prediction of the current noise template depending on the position and velocity of the joints that contribute to the noise generation. Whereas it is prone to modeling errors, it suffers from musical noise components caused by subtraction in the spectral domain. Therefore, multicondition training of acoustic models is not always effective with spectral subtraction based methods, because most speech enhancement techniques distort the spectrum and degrade features. Though the audio signals may be perceived to be cleaner, it does not necessarily mean that the recognition rate is improved. Moreover, template subtraction requires a long training session to build a database of templates to choose from (For more details address to [6]).

Fig. 5 illustrates the ASR accuracy for head motion noise. The results of single-channel MCRA-based background noise reduction are poor, because the level of background noise is considerably lower than the motor noise. Not surprisingly, we observed that GSS+PF operations demonstrate far worse performance compared to GSS alone. That is because short range reverberation effects and multipath propagation are properties of head-motion noise that are very hard to overcome with the current post filter algorithm assumptions and settings. However, we clearly see that only GSS has performed promising results to deal with highly non-stationary head motor noise. For a suitable threshold T, it yields 15% improvement for low SNR's, whereas WERs suffer a considerable reduction when SNR gets higher. We include "best case scenario with known source" for GSS by giving the position of the sound source in advance, which enables us to cross-check the significance of the source separation approach for ego-noise suppression problem. The decrease in the WER's even for high SNR rates ($<$ 20% compared to SSL-dependent GSS approach) prove that a substantial improvement can be achieved in case we can gather correct positions of the sources.

For the second part of the experiment, we recorded head motion noise by rotating the head of ASIMO ($elevation$=$[-30^\circ \ 30^\circ]$, $azimuth$=$[-90^\circ \ 90^\circ]$) randomly. Status information (positions and velocity) of the motors are gathered from the joints with an average acquisition rate of 7.3 ms, slightly faster than our frame shift rate of 10 ms. The training data was a joint database consisting of 30 minutes of motor noise and the corresponding joint-status vectors stored during this time span. We stored a test database of 10 minutes long. In Fig. 5, *TS* indicates template subtraction and *set* specifies the database the templates

are extracted from. *Training set* corresponds to the real experimental condition. *Test set* indicates the usage of *ideal template* constructed from the test set which yields the maximum achievable results for the single-channel approach in that sense. Although the potential of this method is very impressive (as pointed out by the curve labeled with "TS on test set"), template subtraction carried out on training set performs only a minor improvement like 5% to 15%.

After analyzing the capabilities of both single-channel and multi-channel approaches extensively, we suggest to embed both methods into a single system and propose to use them interchangeably in a motion and SNR-specific fashion. Because we can gather information about all active joints and estimated SNR at every time instance, we can apply a switching mechanism between single-channel template subtraction and multi-channel noise reduction methods (See Fig. 1). This switch is triggered by the motion detector's output. Because multi-channel approach works very well for the leg and arm noises, the switch feeds the outputs of this branch to the ASR. On the other hand, in case of a head motion, template subtraction provides more reliable features for high SNR case. If the SNR is low, the switch can either select the multi-channel output or ignore all incoming features depending on the application specifications and confidence requirements of the task.

V. Summary and Outlook

In this paper we presented methods for eliminating ego-motion noise from speech signals. The system we proposed utilizes sound source localization incorporating MUSIC algorithm, sound source separation with GSS algorithm and consequently, speech enhancement stage that suppresses both background noise and interference/leakage noise. We validated the applicability of our approach by evaluating its performance on 3 different motor noise types. Our method demonstrated excellent performance on arm and leg-motion noise. Furthermore, promising results have been presented for the head-motion noise, which is the most challenging type of ego-motion noise due to its close distance to the microphones. To overcome the difficulty of head-motion noise, we proposed to use a hybrid noise reduction system that also incorporates single-channel template subtraction technique in addition to multi-channel approach.

Our system is still open for improvements. One weakness of the current architecture is the threshold value used in the sound source localization procedure, which determines if a source exists at that location. Especially, the higher the motor noise gets, the more susceptible success rates of the system get to the threshold value. There is no optimal threshold value that is effective for every kind of motor noise. Therefore, we plan to make it adaptive. Besides, methods that make use of correlation matrices derived from noise sources in advance, can be very helpful to suppress noise onsets, thus allowing more precise speaker location prediction, causing better separation and higher ASR rates. This system is also capable of dealing with multiple speakers with its current form. Next step is evaluation of the hybrid system in real situation which involves speech recognition of several speakers simultaneously while the robot is performing some task or action.

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", *Proc. of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, 2006.

[2] M. Nakano, A. Hoshino, J. Takeuchi, Y. Hasegawa, T. Torii, K. Nakadai, K. Kato and H. Tsujino,"A Robot that Can Engage in Both Task-oriented and Non-task-oriented Dialogues", *Humanoids*, pp.404-411, 2006.

[3] K. Nakadai, H.G. Okuno, H. Kitano, "Humanoid Active Audition System Improved by The Cover Acoustics", *PRICAI 2000 Topics in Artificial Intelligence (Sixth Pacific Rim International Conference on Artificial Intelligence)*, 544–554, Springer Lecture Notes in Artificial Intelligence No. 1886, 2000.

[4] 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", *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition*, 2006.

[5] A. Ito, T. Kanayama, M. Suzuki, S. Makino, "Internal Noise Suppression for Speech Recognition by Small Robots", *Interspeech 2005*, pp.2685-2688, 2005.

[6] G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino, and J. Imura "Ego Noise Suppression of a Robot Using Template Subtraction", *Proc. of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, pp.199-204, 2009.

[7] S. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, No.2, 1979.

[8] I. Cohen, "Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement", *IEEE Signal Processing Letters*, vol. 9, No.1, 2002.

[9] 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", *Proc. of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, 2006.

[10] L. C. Parra and C. V. Alvino, "Geometric Source Separation: Merging Convolutive Source Separation with Geometric Beamforming", *IEEE Trans. Speech Audio Process.*, vol. 10, No.6, pp. 352-362, 2002.

[11] R. Schmidt, "Multiple emitter location and signal parameter estimation", *IEEE Trans. on Antennas and Propagation*, vol. 34, No.3, pp. 276-280, 1986.

[12] J.-M. Valin, J. Rouat and F. Michaud, "Enhanced Robot Audition Based on Microphone Array Source Separation with Post-Filter", *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2123-2128, 2004.

[13] H. Nakajima, K. Nakadai, Y. Hasegawa and H. Tsujino, "Adaptive step-size parameter control for real-world blind source separation", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.149-152, 2008.

[14] H. Nakajima, K. Nakadai, Y. Hasegawa and H. Tsujino, "High performance sound source separation adaptable to environmental changes for robot audition", *Proc. of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, 2008.

[15] K. Nakadai, H. Nakajima, Y. Hasegawa and H. Tsujino, "Sound source separation of moving speakers for robot audition", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.3685-3688, 2009.

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

[17] I. Cohen and B. Berdugo, "Microphone array post-filtering for non-stationary noise suppression", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.901-904, 2002.

[18] K. Nakadai, H. Okuno, H. Nakajima, Y. Hasegawa and H. Tsujino, "An open source software system for robot audition HARK and its evaluation", *Proc. IEEE-RAS International Conference on Humanoid Robots*, pp.561-566, 2008.