2012 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 7-12, 2012. Vilamoura, Algarve, Portugal

Online Learning for Template-based Multi-channel Ego Noise Estimation

Gökhan Ince, Kazuhiro Nakadai, and Keisuke Nakamura

Abstract— **This paper presents a system that gives a robot the ability to diminish its own disturbing noise (*i.e.,* ego noise) by utilizing template-based ego noise estimation, an algorithm previously developed by the authors. In pursuit of an autonomous, online and adaptive template learning system in this work, we specifically focus on eliminating the requirement of an offline training session performed in advance to build the essential templates, which represent the ego noise. The idea of discriminating ego noise from all other sound sources in the environment enables the robot to learn the templates online without requiring any prior information. Based on the directionality/diffuseness of the sound sources, the robot can easily decide whether the template should be discarded because it is corrupted by external noises, or it should be inserted into the database because the template consists of pure ego noise only. Furthermore, we aim to update the template database optimally by introducing an additional time-variant forgetting factor parameter, which provides a balance between adaptivity and stability of the learning process automatically. Moreover, we enhanced the single-channel noise estimation system to be compatible with the multi-channel robot audition framework so that ego noise can be eliminated from all signals stemming from multiple sound sources respectively. We demonstrate that the proposed system allows the robot to have the ability of online template learning as well as a high performance of noise estimation and suppression for multiple sound sources.**

I. Introduction

The main necessities of mobile robots are to obtain information about the environment by their own and to move their bodies to accomplish useful tasks. Autonomous and online learning of new capabilities/models is their ultimate target because 1) these capabilities/models are too complex to hand-code, 2) they can change in real-time and real-world conditions and, 3) human involved teaching becomes costly both in time and money. Teaching the robot its ego noise, the own noise generated by its motors and fans, is an example of such a task suffering from the above mentioned constraints.

Although general research about the auditory (diffuse/directional) noise estimation techniques date back to a few decades ago, ego noise estimation just recently started to attract interest due to the emergence of robot audition. It has been shown that the estimation of ego noise is crucial in suppressing this noise and consequently achieving good performance in various applications of robot audition such as, Automatic Speech Recognition (ASR) [1] and Sound Source Localization (SSL) [2]. The main difficulty of dealing with ego noise is that firstly the acoustic properties of the motor noise such as the power and frequency characteristics in the spectrum are similar to the properties of sounds of interest such as music or speech, and secondly ego noise is a near-field signal. Therefore, standard solutions designed for far-field signal processing do not work adequately. Moreover, the locations and number of the active motors cannot be assumed to be fixed.

Sawada *et al.* [3] used semi-blind signal separation to obtain ego noise estimates by attaching noise sensors, e.g. Non-Audible Murmur (NAM) microphones inside the robot, but this method requires additional hardware to be mounted inside the robot. Conventional noise estimation techniques [4], [5] fail in estimating the non-stationary ego noise because they are neither able to discriminate ego-motion noise from non-stationary speech signals, nor fast enough to track the rapid changes in ego noise. In contrast to stationary noise estimation methods, template estimation is a better suited method because it represents the motion data using a sequence of observations. Based on these observations, it is possible to associate either a motion command [6] or discrete time series data representing the angular state of motors [1], [7] with another discrete time series data representing the ego noise spectra in the form of templates. The learned associations can then be used to predict an arbitrary sequence of associated data. The so-called Template-based Estimation (TE) has multiple advantages over the conventional stationary noise estimators such that they are not dependent on Signal-To-Noise Ratio (SNR), not prone to Voice Activity Detection (VAD) errors, and the adaptation latency to the actual noise is theoretically zero. The drawbacks of standard TE [1], [6], [7], *i.e.*, constantly growing size of the template database, inability to update the database autonomously and incapacity of coping with changing environmental noise portion in the recorded templates are tackled in our prior work [8]. However, one huge problem still remained, that is how to eliminate the offline training session and human involvement completely.

The learning process, which continues over the entire lifespan of a robot without human intervention, is called life-long learning and it is successfully applied for various tasks such as robot navigation/manipulation [9] and object recognition/categorization [10] in robotics. In this paper we pursue to endow a humanoid robot the ability of life-long learning of templates for template-based ego noise estimation by introducing online learning. The robot learns and updates the templates completely autonomously and in an adaptive manner. This online learning algorithm is a natural extension of the *incremental learning* [8] method and allows the robot

Gökhan Ince, Kazuhiro Nakadai and Keisuke Nakamura are with Honda Research Institute Japan Co., Ltd. 8-1 Honcho, Wako-shi, Saitama 351-0188, Japan {gokhan.ince, nakadai, keisuke@jp.honda-ri.com}

to learn its ego noise not only in isolated and prepared training conditions, but also in daily environments even in the presence of other noise sources such as humans.

The first contribution of the paper is the introduction of the idea of distinguishing ego noise from other types of sound sources using its directivity properties. Based on the instantaneous multi-channel audio spectrum, the robot can easily decide whether the template should be discarded because it is corrupted by external noises, or it should be inserted into the database because the template consists of pure ego noise. Secondly, we aim to update the template database optimally by introducing a time-variant forgetting factor, which provides a balance between adaptivity and stability of the learning process. In contrary to previous template-based suppression methods [1], [6], [7], which can be only applied to a single sound source, we thirdly propose a system, which is able to separate the estimated ego noise contaminating each individual sound source. This novel approach is used to suppress the ego noise on multiple sound sources. Finally, we demonstrate that the proposed system allows the robot to have the ability of online learning as well as improved noise estimation/suppression quality for multiple sound sources.

II. Ego Noise Estimation

In this section we outline the basic architecture of the ego noise estimation system as the flowchart in Fig. 1 depicts. We focus on the major points only. For a more detailed description the reader is advised to consult [1], [8].

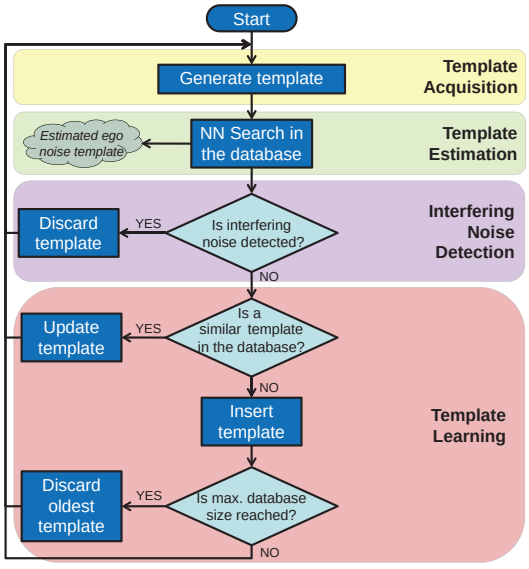


Fig. 1. Flowchart of the proposed ego noise estimation system

A. Template Acquisition

In our prior work [1], we designed a template generation method (the first block in Fig. 1). It utilizes encoders of the J number of joints, which measure the angular position ($\theta(l)$) and velocity ($\dot{\theta}(l)$) of each joint in every frame, l. The resulting feature vector of the template has the form of $\overrightarrow{F}(l) = [\theta_1(l), \dot{\theta}_1(l), \ldots, \theta_j(l), \dot{\theta}_j(l), \ldots, \theta_J(l), \dot{\theta}_J(l)]$. The feature vector is assigned to the ego noise spectral energy vector denoted as $\overrightarrow{N}(l)$ with $N(k,l,m)$ indicating the spectro-temporal energy in frequency bin k, time frame l, and the microphone (channel) m. This data block, $\overrightarrow{T}(l) = [\overrightarrow{F}(l) : \overrightarrow{N}(l)]$, is called a *parameterized template*. In contrary to [1], which was based on a single-channel data, having multiple channels ($m > 1$) allows us to separate ego noise with respect to the sound source it contaminates (cf. Sec. III).

B. Template Estimation

The core of this block is an instance-based, non-parametric classification technique known as the Nearest Neighbor (NN) algorithm. Basically, the spectral energy vector $\overrightarrow{N}^x$ stored in the template $\overrightarrow{T}^x$ of a large template database $\boldsymbol{T}$ with $\overrightarrow{F}^x$ having the shortest distance to $\overrightarrow{F}(l)$ and x being the index of the x-th element in the database is selected as the ego noise estimate $\overrightarrow{\hat{N}}(l)$ [1]. How the template database $\boldsymbol{T}$ is created/updated is the function of the fourth block, template learning, and will be explained in Sec. II-D in detail.

C. Interfering Noise Detection

This block determines whether the current audio signal is corrupted by external noise sources, *i.e.*, any sound except ego noise, such as music, speech, or the extracted template belongs to a pure ego noise signal. Keeping in mind that the ego noise behaves rather like a diffuse signal in the near-field, we put the interfering noise sources into two categories: *directional* and *diffuse* noise sources.

To detect the former type of sources, we propose to use a multi-channel sound source localization method. In order to predict the Directions of Arrival (DoA) of sound sources, we use a popular adaptive beamforming algorithm called MUltiple Signal Classification (MUSIC) [11]. It detects the DoA by performing eigenvalue decomposition on the correlation matrix of the noisy signal. It separates subspaces of undesired interfering sources and sound sources of interest, and finds the peaks in the spatial spectrum. A consequent source tracker system performs a temporal integration in a given time window and a final thresholding is applied to determine the position and (importantly to us) the existence of the directional sound source, if any.

If the noise is not a directional noise such as background noise, MUSIC is unable to detect it, but the explicit detection of the diffuse noise is not a crucial issue to the overall estimation system at all because the templates already contain the accumulated diffuse noise of background and ego noise. They can adapt to the changes in the diffuse noise by using the online template learning method presented in the next section. Thus, from now on, "interfering noises" will refer to the directional noise sources.

D. Template Learning

The main goal of the last block of Fig. 1, as proposed in [8], is to learn templates incrementally and include them to the template database. For this purpose, the output of the NN algorithm in the template estimation block can be interpreted as the similarity of the input pattern to the estimated pattern. This measure, also known as the relative confidence level,

allows us to determine if each observed template is a previously known template or a new template to be learned. Based on the comparison of a given fixed distance threshold, d_t, with $d_{min}(\overrightarrow{F}(l), \overrightarrow{F}^x)$ having the smallest vectoral distance between the current template $\overrightarrow{F}(l)$, and x-th vector $\overrightarrow{F}^x$ in $\boldsymbol{T}$, the current template is either used to update the old template or it is inserted into the database as a new template. When the similarity is low, the template is treated as a missing template and inserted into $\boldsymbol{T}$; otherwise the adaptive update mechanism is active. In conjunction with the interfering noise detection block, this block builds up an autonomous and online learning system.

In [8], the contribution of past templates were reduced by introducing a fixed forgetting factor ($\eta = 0.9$), which computed the weighted average of the old and current template by laying the focus more on recently-acquired templates and less on earlier observations.

We introduce a time-variant forgetting factor to enhance the balance between adaptivity (learning quality) and stability (robustness against errors and unexpected transient noises, *e.g.*, mechanical jittering and shuddering sounds). The former is achieved by using lower $\eta(l)$, whereas higher $\eta(l)$ enables stability. Its computation is as follows:

$$\eta(l) = \frac{\frac{1}{1+\exp(-\sigma_1 d_{min}(\overrightarrow{F}(l), \overrightarrow{F}^x))} + \frac{1}{1+\exp(-\sigma_2 \varepsilon(l))}}{2}, \quad (1)$$

$$\varepsilon(l) = \frac{\sum_{k=0}^{K} ||N(k,l,1)|^2 - |\hat{N}(k,l,1)|^2|)}{\sum_{k=0}^{K} ||N(k,l,1)|^2|}, \quad (2)$$

where $\varepsilon(l)$ is the normalized noise estimation error of the first microphone signal, σ_1 and σ_2 are tilt values for the sigmoid functions. To reduce the computational cost, we selected $m = 1$ as a single representation of all channels. So, $\eta(l)$ takes values between 0.5 and 1. When either the estimation error regarding the features, $d_{min}(l)$, or regarding the spectra, $\varepsilon(l)$, is large, $\eta(l)$ increases. As a consequence, the contribution of the new erroneous template is reduced. The pseudo-code of the final online learning algorithm is shown below.

Algorithm 1 Learning of templates

if $d_{min}(\overrightarrow{F}(l), \overrightarrow{F}^x) \geq d_t$ **then**
  $[F^{new}(j,l) : N^{new}(k,l,m)] \leftarrow [F^{curr}(j,l) : N^{curr}(k,l,m)]$
else
  $[F^{old}(j,l) : N^{upd}(k,l,m)] \leftarrow [F^{curr}(j,l) : \eta(l)N^{old}(k,l,m) + (1-\eta(l))N^{curr}(k,l,m)]$
end if

This block also checks if the requirement of a limited database size is violated, thus it makes sure that the size limitations are not exceeded.

III. System Architecture

The proposed multi-channel noise reduction system is illustrated in Fig. 2. The audio signals are firstly subject to SSL modules and then to the ego noise estimation modules. As explained in Sec. II-C, the output of SSL is interpreted as a trigger in the template learning module to decide whether to apply learning or discarding process. Also, the source locations constitute the input of Sound Source Separation (SSS) along with the audio spectra to separate the useful audio signal, as well as of a second SSS module along with the estimated multi-channel noise template to separate the overall ego noise among all sound sources. By doing so, spectral subtraction is applied on the spectrum of each individual sound source using its corresponding ego noise spectrum.

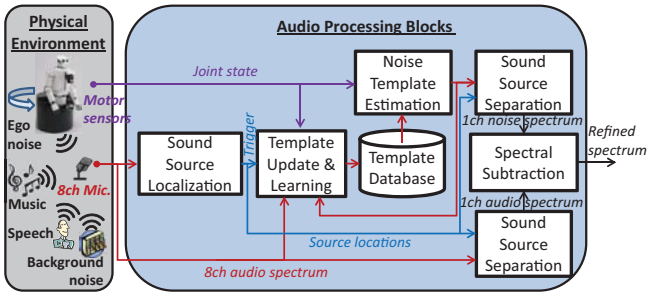
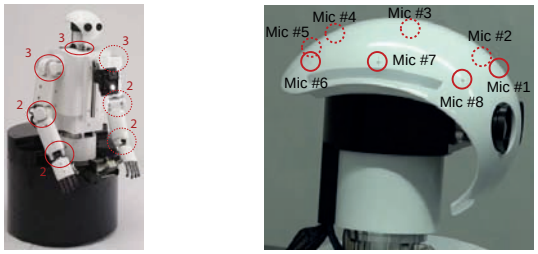


Fig. 2. Block diagram of the proposed noise suppression system

IV. Evaluation

A. Experimental Settings

1) Hardware Specifications: The used robotic platform was a humanoid robot called HEARBO (Fig. 3(a)). We used its 17 Degrees of Freedom, and an 8 channel omni-directional microphone array on top of its head, as illustrated in Fig. 3(b). The audio signals were synchronously captured by a RASP-24 bits unit at a 16 kHz sampling rate, and transmitted at 100.0 Hz. The joint sensor data was acquired at 100.0 Hz as well. All processes were handled by an Intel Core i5 quadcore laptop PC at 2.53 GHz, with 8 GB of RAM.



(a) Positions and number of moving joints (b) Close-up of the head

Fig. 3. HRI-JP humanoid robot HEARBO

2) Software Specifications: The audio spectrum was computed using a Complex window with 512 samples with a hop size of 160 samples. For SSS, we used Geometric High-order Decorrelation-based Source Separation (GHDSS) [12] algorithm. The size limit of the template database is set to 10000, which was never reached in our experiments. The whole signal processing architecture was implemented on *HARK*[1], an open-source software for robot audition. The motion generation, recording processes and the bi-directional dataflow between HARK and the robotic platform were handled by *ROS*[2].

[1]http://winnie.kuis.kyoto-u.ac.jp/HARK/
[2]http://www.ros.org

3) Experimental Setup: For the recordings and processing we used 8 channel microphones, but for the comparison with single-channel results only the frontal microphone #1 of the robot (see Fig. 3(b)) was used. The recordings were performed in a noisy room with the dimensions of 4.0 m×7.0 m×3.0 m with a Reverberation Time (RT_{20}) of 0.2 sec. We generated a training and a test set of periodic motions consisting of predefined behaviors. The names of the motions and the DoFs actively involved (L: Left arm, R: Right arm, H: Head) are enlisted in Tab. I. To study the performance of the proposed system and be able to specifically assess the detection and estimation quality, we needed to control the conditions. Therefore, we mixed the interfering noise sources, *i.e.* speech and music signals, which were convoluted with the impulse response of the experimental environment, with the recorded ego noise signals under varying SNR conditions. The ego noise files were 60 sec long and contained multiple iterations of the same motions. The interfering noise files were 20 sec long. The interfering noises were mixed exactly in the middle of the ego noise (between 20-40 sec). The evaluation of our system for uncontrolled real-world conditions are presented in our complementary paper [13].

B. System Evaluation

In this section, we describe the assessment methods for each important aspect of the proposed system, namely *1)* the success rate of distinguishing ego noise from other noise sources, *2)* the quality of the estimation using the time-variant forgetting factor and online system, and *3)* the performance of the suppression using the multi-channel approach.

1) Interfering Noise Detection: We compared the performance of the proposed multi-channel SSL method (M1) with two other single-channel noise detection algorithms that use energy-based thresholding strategies. We selected the thresholds by *i)* computing the average energy from a template database collected *a priori* (M2), and *ii)* performing a weighted k-means clustering on the features of the same database (M3). Whereas the former approach (M2) was rather brute force because a fixed threshold was used for a wide variety of motions with different energies, the latter approach (M3) was more reasonable because it took the similarity of the features and their corresponding energies inside each cluster into account. Based on our preliminary experiments, we empirically weighted the velocities 5 times more than the positions in the feature vectors because essentially the higher the velocities the higher is the ego noise energy. We also empirically created $k = 50$ clusters. Then, we computed a set of thresholds for each cluster by calculating the average energy of all templates belonging to each cluster. In the estimation phase, the algorithm of M3 selected the corresponding threshold for each instantaneous audio frame of ego noise based on the similarity of the current joint status to the cluster centers. Both methods (M2 and M3) were able to adapt their thresholds in the run-time. Nevertheless, they required a small prior template database to build the initial knowledge, which also contradicts the elimination of the offline session requirement for online learning. We also tested two databases for M2 and M3; one acquired from the same motions as in test motions and one from arbitrary motions.

2) Noise Estimation: Incremental learning [8] made it possible to correct the errors in the training set; additionally, the proposed online system with time-variant forgetting factor ($\eta(l)$) enables the system to quickly and optimally adapt to partially-known or dynamically changing environments. In [8], the "long-term" estimation performance after numerous iterations of the same motion has been already elaborately assessed for the offline incremental system in relation with the threshold d_{min} and database size. Thus, in this work we rather assess the "short-term" performance of the noise estimation using the online learning within only a few iterations, in other words the applicability of the online system in realistic conditions.

Because the interfering noises, *i.e.* speech and music, were mixed in the middle of the ego noise (between 20-40 sec), the system uses the first 20 sec to adapt to the motion. Taking into account that all tested motion instances last 5-20 sec, the online system has sufficient but not extensive time to create a reasonable database. The mixed audio signal between 20-40 sec is used for the evaluation of the estimation on the fly. We used empirical values for $\sigma_1 = 10$ and $\sigma_2 = 10$.

3) Noise Suppression: We used spectral subtraction using a *single-channel* of the templates corresponding to microphone #1 and the whole *multi-channel* templates to compare the performance of the previous and proposed system. In both cases, we used a minor spectral floor of 0.1. The assessment was undertaken using an ASR task.

C. Evaluation Criteria

This section explains the criteria used to assess each aspect of the proposed system as described in Sec. IV-B.

1) F-measure and Accuracy: To assess the performance of the interfering noise detection, we use common statistical analysis criteria for detection tasks called *F-Measure* and *accuracy*. They are computed using *True Positive (TP)*, *True Negative (TN)*, *False Positive (FP)*, *False Negative (FN)*:

$$Precision = \frac{TP}{TP+FP}. \tag{3}$$

$$Recall = \frac{TP}{TP+FN}. \tag{4}$$

$$F-Measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}. \tag{5}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}. \tag{6}$$

2) Normalized Noise Estimation Error (NNEE): NNEE computes the error of the noise estimate normalized by the energy of the actual noise using the following formula:

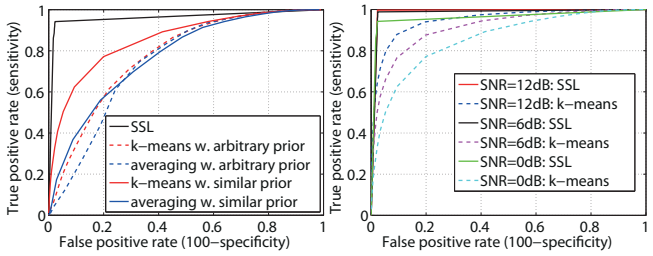
$$\bar{\varepsilon} = \frac{1}{L}\sum_{l=1} 10 \cdot \log_{10}\left(\frac{\sum_{k=1,m=1}^{K,M} ||N(k,l,m)|^2 - |\hat{N}(k,l,m)|^2|}{\sum_{k=1,m=1}^{K,M} ||N(k,l,m)|^2|}\right), \tag{7}$$

where L is the number of frames, and M is the number of microphones.

3) Automatic Speech Recognition: The noise signals are mixed with clean speech utterances used in a typical human-robot interaction dialog and recorded by us. This a Japanese word dataset includes 236 words for 4 female and 4 male speakers. We used matched acoustic models 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 and speaker-open test. For ASR, we used 13 static Mel-Scale Log Spectrum (MSLS) features, 13 delta MSLS features and 1 delta power feature. Speech recognition results are given as average Word Correct Rates (WCR) of instances from the noisy test set.

D. Results

1) Interfering Noise Detection Performance: Fig. 4(a) shows the results using Receiver Operating Characteristic (ROC) curves for our binary detection system averaged for all motions and interfering noises as the thresholds are varied for multi-channel SSL and single-channel clustering-based methods ($SNR = 0dB$). First of all, we see that the performance of single-channel thresholding is better when the thresholds are trained using a database consisting of the similar motions, when compared to a database with arbitrary motions for both M2 and M3. In other words, the former database provides the best case scenario that the adaptive template learning system can achieve after several iterations. Secondly, the clustering-based detector (M3) is better-suited to detect the interfering noise compared to a simple averaging-based detector (M2), but both of these methods are inferior to SSL-based detector (M1) by far. The final observation is that the higher the SNR, the more accurate is the detector (see Fig. 4(b)).



(a) ROC for different methods (b) ROC for different SNRs

Fig. 4. ROC curves of interfering noise detection. M1:SSL, M2:k-means, M3:averaging

The strength of the multi-channel approach is that it utilizes the propagation characteristics of the noise sources, taking advantage of the ego noise being rather diffuse and the interfering noise sources being highly directional. Fig. 5(a) shows a spatio-temporal graph of the *dance* motion noise using SSL, Fig. 5(b) shows the graph of a music signal as an interfering noise and Fig. 5(c) depicts how the directivity is still detectable in the noisy mixture even in a low SNR of 0dB. On the other hand, the single-channel approach using thresholding is solely based on the spectral energy characteristics of the signals, which is problematic because the ego noise is difficult to be distinguished from speech or music signals in terms of spectral energy as can be seen in Fig. 6. Thus, in the remainder of this paper we will use SSL-based interfering noise detection to train the templates.

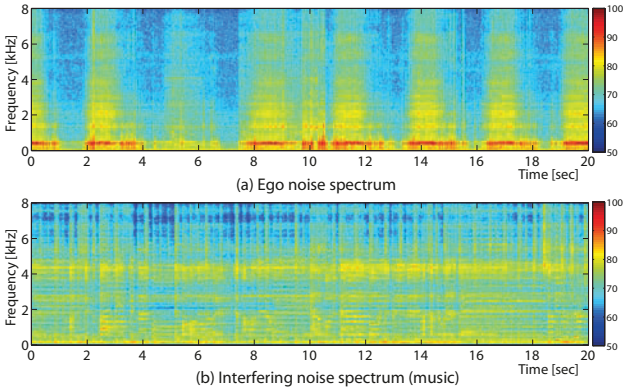


Fig. 6. Spectra of the signals between [20-40] sec of Fig. 5

Tab. I shows the final performance in terms of accuracy and F-measure given a threshold value of 34 is used in the SSL (see Fig. 5(c)). We also see that the rates are mostly correlated to the noise level of the motions.

TABLE I
INTERFERING NOISE DETECTION PERFORMANCE

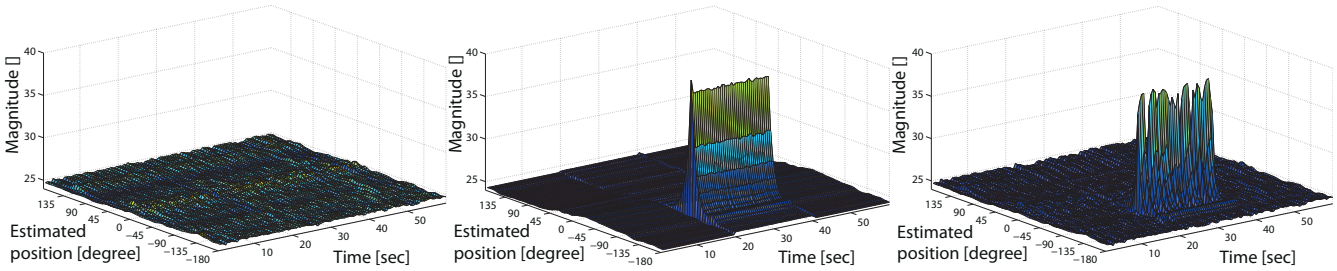
Motion	*Joints*	*SNR[dB]*	*Accuracy*	*F-Measure*
Banzai	L,R	-3.2	0.89	0.82
Bow	L,R,H	2.9	0.98	0.97
Confused	L,R,H	4.1	0.97	0.96
Dance	L,R	-5.1	0.95	0.93
Here	L,R	0.7	0.98	0.97
No	H	-2.9	0.94	0.91
Point Back	L,R,H	-1.0	0.96	0.94
Start	L,R	5.6	0.98	0.98
Wave	L,R	2.5	0.98	0.97
Yes	H	-0.4	0.98	0.97
Overall	-	-	**0.96**	**0.94**

2) Noise Estimation Performance: Tab. II shows the comparison of the estimation performance when we use the following three techniques: (1) continuous insertion of the templates to the database [1], (2) incremental learning of templates with $\eta = 0.9$ [8], and (3) incremental learning of templates with adaptive $\eta(l)$. "Offline" indicates that the database was previously created from a separate training data of ego noise for 60 sec, whereas "online" uses the first 20 sec of the test data to build the database and evaluates it with noisy data in the next 20 sec autonomously.

TABLE II
EGO NOISE REDUCTION PERFORMANCE FOR ALL METHODS

Learning Method	*Offline*	*Online*
Continuous insertion [1]	-4.43±0.63	-4.14±0.66
Fixed incremental learning [8]	-4.85±0.41	-4.70±0.26
Adaptive incremental learning	-5.02±0.35	-4.88±0.30

We see clearly that the adaptive incremental learning improved the performance compared to our previous two

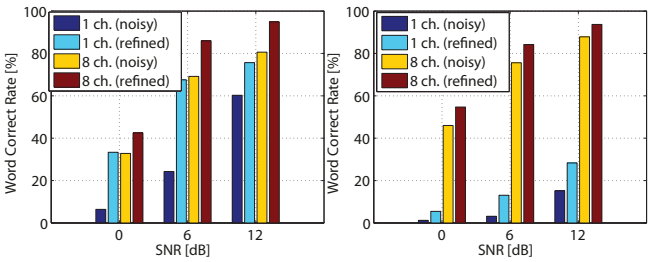


(a) Ego noise (dance motion) (b) Interfering noise (music) from −90° (c) Noisy signal (ego noise+music)

Fig. 5. MUSIC spectrum in each frame during a dance motion

approaches. Furthermore, an online learning system utilizing an interfering noise detection mechanism worked almost as good as the offline system, which shows the validity and applicability of the online approach.

3) Noise Suppression Performance: We evaluated the ASR performance of the online system in two conditions: (a) Single speaker talking from 60° while the robot was moving, and (b) Single speaker talking from 60° while music was being played from −60° and the robot was moving. The average results over all motions are shown in Fig. 7. *1ch (noisy)* and *8ch (noisy)* indicate speech signals extracted from single-channel and GHDSS applied to multi-channel audio data, whereas *1ch (refined)* and *8ch (refined)* show the same signals after ego noise suppression, respectively. We see that the performance of single-channel noise suppression is as good as GHDSS in (a), but the existence of the second sound source makes it impossible for it to operate in the environments with multiple sound sources such as (b). However, the proposed multi-channel ego noise suppression system improves WCRs in both conditions considerably.



(a) WCR for single speaker (b) WCR for single speaker+music

Fig. 7. ASR results

V. Summary and Outlook

In this paper we proposed an online learning mechanism for a multi-channel template-based ego noise estimation framework, which is able to suppress the ego noise contaminating multiple sound sources. We assessed the learning, estimation and suppression performance of this noise reduction method in the presence of multiple types of ego-motion noise. The proposed time-variant forgetting factor decreased the estimation error. We also showed that the interfering noise detection eliminates offline training sessions and enables online learning, which still attains precise estimation of overall noise and high ASR accuracy under various SNR conditions.

To achieve *optimal* performance, we suggest to keep the learning/updating algorithm passively running as a background process for *longer periods* while the robot is performing its tasks or rehearsing the motions, without taking the external noises into account. Future plans include integrating a stationary noise estimator prior to ego noise estimation, which is found to be useful for improving the overall noise estimation results as suggested in [8].

References

[1] G. Ince, K. Nakadai, T. Rodemann, H. Tsujino, and J. Imura, "Whole body motion noise cancellation of a robot for improved automatic speech recognition", *Advanced Robotics*, vol. 25, no. 11, pp. 1405-1426, 2011.

[2] G. Ince, K. Nakamura, F. Asano, H. Nakajima and K. Nakadai, "Assessment of general applicability of ego noise estimation - application to automatic speech recognition and sound source localization-", *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3517-3522, 2011.

[3] H. Sawada, J. Even, H. Saruwatari, K. Shikano, T. Takatani, "Improvement of speech recognition performance for spoken-oriented robot dialog system using end-fire array", *Proc. of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, pp. 970-975, 2010.

[4] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments", *Signal Processing*, vol 81, pp. 2403-2481, 2001.

[5] H. Nakajima, G. Ince, K. Nakadai and Y. Hasegawa, "An easily-configurable robot audition system using histogram-based recursive level estimation", *Proc. of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, pp. 958-963, 2010.

[6] Y. Nishimura, M. Ishizuka, K. Nakadai, M. Nakano, and H. Tsujino, "Speech recognition for a robot under its motor noises by selective application of missing feature theory and MLLR", *Proc. of the IEEE-RAS International Conference on Humanoid Robots*, pp. 26-33, 2006.

[7] A. Ito, T. Kanayama, M. Suzuki, S. Makino, "Internal noise suppression for speech recognition by small robots", *Proc. of the Interspeech 2005*, pp. 2685-2688, 2005.

[8] G. Ince, K. Nakadai, T. Rodemann, J. Imura, K. Nakamura and H. Nakajima "Incremental learning for ego noise estimation of a robot, *Proc. of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, pp. 131-136, 2011.

[9] S. Thrun, *Explanation-based neural network learning: a lifelong learning approach*, Kluwer Academic, 1996.

[10] S. Kirstein, H. Wersing, E. Körner, "Towards autonomous bootstrapping for life-long learning categorization tasks", *Proc. of the International Joint Conference on Neural Networks*, pp. 1-8, 2010.

[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] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino. "Blind source separation with prameter-free adaptive step-size method for robot audition", *IEEE Trans. Audio, Speech, and Language Processing*, 18(6):14761484, 2010.

[13] J. L. Oliveira, G. Ince, K. Nakamura, K. Nakadai, H. G. Okuno, L. P. Reis, F. Gouyon, "Live assessment of beat tracking for robot audition", submitted to *IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, 2012.