2011 IEEE International Conference on Robotics and Automation
Shanghai International Conference Center
May 9-13, 2011, Shanghai, China

Assessment of General Applicability of Ego Noise Estimation
– Applications to Automatic Speech Recognition and Sound Source Localization –

Gökhan Ince, Keisuke Nakamura, Futoshi Asano, Hirofumi Nakajima and Kazuhiro Nakadai

Abstract— **Noise generated due to the motion of a robot deteriorates the quality of the desired sounds recorded by robot-embedded microphones. On top of that, a moving robot is also vulnerable to its loud fan noise that changes its orientation relative to the moving limbs where the microphones are mounted on. To tackle the non-stationary ego-motion noise and the direction changes of fan noise, we propose an estimation method based on instantaneous prediction of ego noise using parameterized templates. We verify the ego noise suppression capability of the proposed estimation method on a humanoid robot by evaluating it on two important applications in the framework of robot audition: (1) automatic speech recognition and (2) sound source localization. We demonstrate that our method improves recognition and localization performance during both head and arm motions considerably.**

I. Introduction

Robots with microphones are usually equipped with adaptive noise cancellation and acoustic echo cancellation methods for robust automatic speech recognition (ASR) and sound source localization (SSL) in noisy environments. However, the robot's own noise, so called ego noise, can cause mis-recognition of spoken words during an interaction with a human, even if there are no other interfering sound sources in an 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 noise gets even more severe for a moving robot with a high degree of freedom, like a humanoid robot. Although the second type of ego noise, the fan noise, is louder, ego-motion noise is more difficult to be coped with, because it is non-stationary and, to a certain extent, similar to the signals of interest in terms of its directivity property [1]. Therefore, conventional noise reduction methods like spectral subtraction [2] do not work well in practice. A directional noise model such as assumed in the case of interfering speakers [3] or a diffuse background noise model [4] do not represent ego-motion noise characteristics entirely either. Especially because the motors are located in the near field of the microphones and are covered with body shells, they emit sounds having both diffuse and directional characteristics, which makes this noise difficult to predict. On the other hand, the noise emitted from the fan of the robot is the main reason of mis-recognition of the sound sources. When the robot moves its limbs on which the microphones are mounted, the direction of ego noise alters rapidly, therefore the effects created by the moving microphones must be taken care of.

Nishimura *et al.* [5] and Ito *et al.* [6] tackled the ego noise problem by predicting and subtracting ego-motion noise using templates recorded in advance for each motion and gesture involving activity of several motors at a time, but their methods work only for a limited number of gestures and motions with fixed trajectories. Even *et al.* [7] proposed to use semi-blind signal separation to obtain both external and internal noise by attaching additional sensors inside the robot. After a Wiener filter-based suppression step, a delay-and-sum beamformer enhances the refined speech. Although it improves speech recognition accuracy considerably, this method requires a body cover made of high-quality or thick material so that the assumption can hold that external noise is definitely not recorded by these additional sensors. Previously, we presented an ego noise estimation method based on instantaneous prediction of ego noise using parameterized templates [9], which can be implemented on any mobile robot regardless of any physical constraint about its external shielding and exploits only existing microphones. An important feature of this method is that it is well-suited to capture the dynamic nature of the motion data represented by the sequence of observations. Based on these observations, we were able to associate a discrete time series data (motion) with another discrete time series data (ego noise) and predict an arbitrary sequence of associated data. We also reported a basic system utilizing this approach to achieve an ASR task during ego motion of a humanoid robot [1]. By exerting Missing Feature Theory (MFT), Yamamoto *et al.* [3] and Takahashi *et al.* [8] proposed models for mask generation to eliminate leakage noise in a simultaneous speech recognition task of several speakers, however their models are unable to deal with ego-motion noise. In a related study, which aims to solve the ego noise problem in a multi-talker ASR application using MFT, Ince *et al.* introduced a masking model based on the Signal to Noise Ratio (SNR) of the ego noise estimates [10]. All these above-mentioned studies focused on ASR, however there is even less work that pursues a robust SSL under ego motion noise. In most studies, either the angular velocity of motors are reduced to create less noise [11], or the sound processing is performed by following *Act-Stop-Sense* principle [12]. Nakadai *et al.* [13] proposed a noise cancellation method with two pairs of microphones. One pair in the inner part

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

Gökhan Ince and Kazuhiro Nakadai 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

978-1-61284-380-3/11/$26.00 ©2011 IEEE

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, but its performance is not satisfactory.

In this paper, we extend the application domain of ego noise estimation to two important processes from the field of ”Robot Audition”, which pursues to achieve general sound understanding: ASR and SSL. The main contributions of our work are (1) further improvement of basic ASR system [1] with adaptive noise superimposition and utilization of Missing Feature Theory (MFT), and (2) application to SSL to demonstrate the general applicability of our ego noise estimation method. Both applications utilize a common ego noise prediction subsystem and a generic subsystem explicitly designed to establish ASR or SSL (See Fig. 2(a) and Fig. 2(b)). For the ASR application, we complement the ego noise estimation system with MFT that applies a filtering operation to the damaged acoustic features that are subject to residuals of motor noise. For the SSL application, ego noise estimation system is used in combination with an SSL system to decorrelate the ego noise and cope with head rotation effects. We show that the proposed methods achieve a high noise elimination performance for both applications.

II. Ego-motion Noise Prediction

The underlying motivation of using templates for noise prediction resides in the fact that the duration and the envelope of the motor noise signals do not change drastically for the same motions when the motion is performed again. A conventional *blockwise template prediction* [5] that extracts templates as a single block has several shortcomings, e.g. it could be performed properly only after the detection of the exact starting moment of the template. Another drawback is that it requires a large collection of data consisting of the motor noise statistics for each joint of different combinations of origin, target, position, velocity and acceleration parameters. To overcome these deficits, we implement *parameterized template prediction* technique [9] that fragments a discrete audio segment into frames by associating them with the current status of the motors. The data is provided by the joint angle sensors that measure the angular positions of all joints separately.

A. Motion Prediction and Template Database Generation

During the motion of the robot, actual position (θ) information regarding each motor is gathered regularly. Using the difference between consecutive sensor outputs, velocity ($\dot{\theta}$) and acceleration ($\ddot{\theta}$) values are calculated. Considering that J joints are active, $3J$ attributes are generated. Each feature is normalized to [-1 1] so that all features have the same contribution on the prediction. The resulting feature vector has the form of $\mathfrak{F}(k) = [\theta_1(k), \dot{\theta}_1(k), \ddot{\theta}_1(k), \ldots, \theta_J(k), \dot{\theta}_J(k), \ddot{\theta}_J(k)]$, where k stands for the time-frame. At the same time, motor noise is recorded. The spectrum of the motor noise is given by $D_n(k) = [D_n(1,k), D_n(2,k), \ldots, D_n(F,k)]$, where ω is discrete frequency, F represents the number of frequency bins and n denotes the index of a microphone. Both feature vectors and spectra are continuously labeled with time tags so that corresponding templates are generated when their time tags match. As will be explained in Sec. III-A and Sec. III-B, the number of simultaneously recorded spectra (n) depends on the requirements of the application.

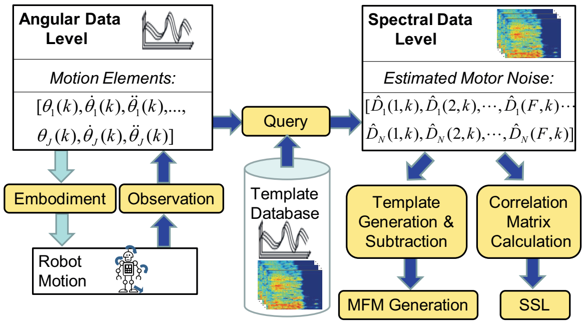

Angular Data Level
Motion Elements:
$[\theta_1(k), \dot{\theta}_1(k), \ddot{\theta}_1(k), \ldots,$
$\theta_J(k), \dot{\theta}_J(k), \ddot{\theta}_J(k)]$
Embodiment
Observation
Robot Motion
Query
Template Database
Spectral Data Level
Estimated Motor Noise:
$[\hat{D}_1(1,k), \hat{D}_1(2,k), \cdots, \hat{D}_1(F,k) \cdots$
$\hat{D}_N(1,k), \hat{D}_N(2,k), \cdots, \hat{D}_N(F,k)]$
Template Generation & Subtraction
Correlation Matrix Calculation
MFM Generation
SSL

Fig. 1. Parameterized template prediction method and its applications

B. Parameterized Template Prediction

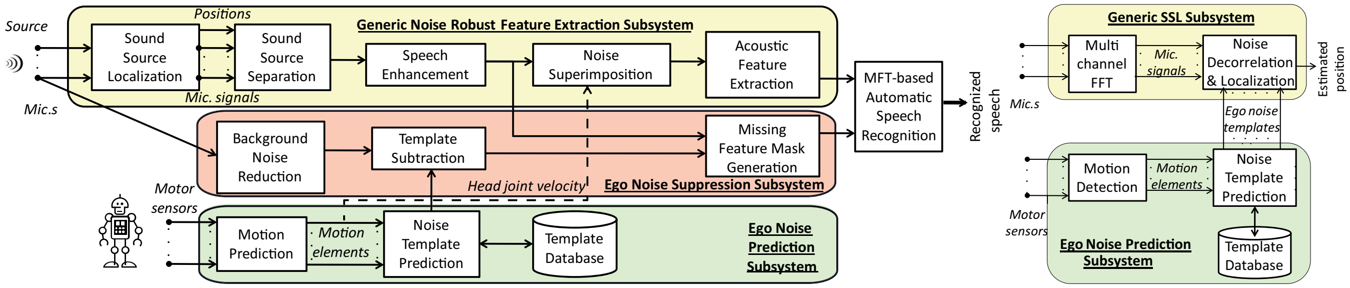
The prediction phase starts with a search in the database for the best matching template of motor noise for the current time instance (See Fig. 1). We implemented a Nearest Neighbor search to find the correct template with the most similar joint configuration among all templates in the database. The prediction process is applied to every frame. In that sense, the conventional *blockwise template* for a single arbitrary motion can be regarded as the concatenation of smaller templates that are predicted according to the above-mentioned approach on a frame-by-frame basis.

III. Applications of Ego Noise Estimation

We investigate the applicability of ego noise estimation (including its extensions such as in Sec.III-A.1–Sec.III-A.3) on two essential robot audition tasks: ASR and SSL.

A. Ego Noise Robust Automatic Speech Recognition

In this section, we describe a standard ASR system using a microphone array, which is robust to environmental noise and interfering speakers (see Fig. 2(a)). The chain starts with an SSL module. In order to estimate the location of the speaker, we use one of the most popular adaptive beamforming algorithms called MUltiple Signal Classification (MUSIC). It detects the locations of sources by performing an eigenvalue decomposition on the correlation matrix of the noisy signal and sends them to the Sound Source Separation (SSS) stage, which is a linear separation algorithm called Geometric Source Separation (GSS) [3]. After the separation process, a multi-channel post-filtering (PF) operation proposed by Cohen [14] is applied, which can cope with stationary noise. Details about the usage of this processing chain can be found in [1]. A consequent additive white noise step improves the speech recognition results by generating an artificial floor in the spectrum of speech signal. Finally, acoustic features are generated by calculating Mel-Scale Log Spectrum (MSLS)


Source
Mic.s
Positions
Mic. signals
Generic Noise Robust Feature Extraction Subsystem
Sound Source Localization
Sound Source Separation
Speech Enhancement
Noise Superimposition
Acoustic Feature Extraction
MFT-based Automatic Speech Recognition
Recognized speech
Background Noise Reduction
Template Subtraction
Missing Feature Mask Generation
Head joint velocity
Ego Noise Suppression Subsystem
Motor sensors
Motion Prediction
Motion elements
Noise Template Prediction
Template Database
Ego Noise Prediction Subsystem
Generic SSL Subsystem
Multi channel FFT
Mic. signals
Noise Decorrelation & Localization
Estimated position
Mic.s
Ego noise templates
Motion Detection
Motion elements
Noise Template Prediction
Motor sensors
Ego Noise Prediction Subsystem
Template Database

(a) Proposed automatic speech recognition system

(b) Proposed SSL system

Fig. 2. Two major applications of ego noise estimation

that maintains distortions in specific spectral bins unlike Mel-Frequency Cepstral Coefficients (MFCC).

1) White Noise Superimposition: Because it is impractical to create matched models for each ego-motion noise, we add white noise with a fixed amplitude value as a known noise source during the training phase. The second advantage of using white noise is that it blurs the *musical noise* [2] distortions caused by the spectral subtraction of the PF. Because the artifacts of the louder motor noise are more harmful compared to the artifacts of less noisier motors, we propose a switching mechanism for white noise level adjustment inside the noise superimposition module. The mechanism performs a decision between two white noise levels $\{C_1, C_2\}$, which is triggered by the motion predictor. This method is scalable according to the physical conditions regarding microphones, motors, their distances and properties. We propose to implement the following rule-based routing in the switch:

$$\rho(k) = \begin{cases} C_1, & \text{if any } |\dot{\theta}_{LoudJoints}(k)| > \varepsilon \\ C_2, & \text{otherwise} \end{cases}, \tag{1}$$

where ρ $[dB]$ represents white noise magnitude relative to clean speech magnitude, $|\dot{\theta}_{LoudJoints}(k)|$ denotes absolute velocity of the related joint and ε is a certain speed value. ε, instead of zero, is used to prevent the activation of the switch when the motion has stopped, but the joint sensors still send very small position differences. Motion detection is compromised by a high ε value. Please note that the additive white noise will be cancelled out in the spectral mean normalization module of ASR.

2) Template Subtraction (TS) [9]: We start by defining $S(\omega,k)$ and $D(\omega,k)$ as the short-time basis spectra of speech signal and distortion (motor noise only), respectively, where ω stands for the discrete frequency representation. So, the spectrum of the observed signal $X(\omega,k)$ can be given as:

$$X(\omega,k) = S(\omega,k) + D(\omega,k). \tag{2}$$

The spectrum of the useful signal can be obtained by using the inverse operation of Eq. (2):

$$S_r(\omega,k) = X(\omega,k) - \hat{D}(\omega,k), \tag{3}$$

where $\hat{D}(\omega,k)$ denotes the estimated noise template and $S_r(\omega,k)$ stands for the signal comprising the useful sound and residual motor noise. The reason of this residual noise is that the original motor noise $D(\omega,k)$ deviates from the predicted one. To compensate for this error, we further suggest to use the spectral subtraction approach that exploits *overestimation factor*, α, and *spectral floor*, β. α, allows a compromise between perceptual signal distortion and the noise reduction level, whereas β is required to deal with musical noise. Finally, we calculate the gain coefficients, $\hat{H}_{SS}(\omega,k)$, and multiply them with the signal $X(\omega,k)$ as in Eq. (5):

$$\hat{H}_{SS}(\omega,k) = max\left(1 - \alpha\frac{|\hat{D}_{(}\omega,k)|}{|X(\omega,k)|}, \beta\right). \tag{4}$$

$$\hat{S}(\omega,k) = \hat{H}_{SS}(\omega,k) \cdot X(\omega,k). \tag{5}$$

It is noteworthy that in contrary to [1] and [9], the templates are subtracted from the noisy signal only to obtain the soft masks and not to suppress the noise directly.

3) Missing Feature Mask (MFM) Generation: The problem with the proposed feature extraction subsystem in Fig. 2(a) is that when the position of the noise source is not detected precisely, SSS cannot separate the sound in the spatial domain precisely as well. As a consequence, motor noise can be spread to the separated sound sources in small portions. However, it is optimally designed for *simultaneous speakers* scenarios with background noise and demonstrates a 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 pre-recorded template of the motor noise at any moment. The drawback of this approach is, however, due to the non-stationarity or missing templates in the database, the predicted and actual noise can differ.

As stated above, the strengths and weaknesses of both approaches are distinct. Thus, they can be integrated into an MFM in a complementary fashion. In that sense, a speech feature can be considered unreliable, if the difference between the energies of refined speech signals generated by multi-channel (SSL+SSS+SE) and single-channel (TS) noise reduction systems is large. Computation of the masks is performed for each frame, k, and for each Mel-frequency band, f. First, a continuous mask is calculated as follows:

$$m(f,k) = 1 - \left(\frac{\left||\hat{S}_m(f,k)|^2 - |\hat{S}_s(f,k)|^2\right|}{|\hat{S}_m(f,k)|^2 + |\hat{S}_s(f,k)|^2}\right), \tag{6}$$

where $|\hat{S}_m(f,k)|^2$ and $|\hat{S}_s(f,k)|^2$ are the estimated energies of the refined speech signals, which were subject to multi-channel noise reduction and resp. single-channel template subtraction. 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.) A soft mask as in Eq.(7) [8] is used in the MFT-ASR to control the sensitivity of $m(f,k)$:

$$M(f,k)=\begin{cases}\dfrac{1}{1+\exp(-\sigma(m(f,k)-T))}, & \text{if } m(f,k)\geq T\\ 0, & \text{if } m(f,k)<T\end{cases}, \tag{7}$$

where σ is the tilt value of a sigmoid weighting function and T represents the threshold.

B. Ego Noise Robust Sound Source Localization

In a robotic system with general audition capabilities, SSL results affect the consequent stages of SSS and ASR implicitly. Therefore, the noise must be suppressed in the spatial domain to achieve sound localization accurately, especially for a dynamical environment with a low signal-to-noise ratio. This section describes an SSL system, which is able to decorrelate the noise from the noisy signal captured by a microphone array (see Fig. 2(b)). For this application, we propose to use MUSIC based on the Generalized Eigen Value Decomposition (GEVD) [15] technique. Contrary to Standard Eigenvalue Decomposition-MUSIC (SEVD-MUSIC), it utilizes a noise correlation matrix in order to suppress environmental noise sources.

Suppose that we have M sources and N ($> M$) microphones. $\mathbf{X}(\omega) = [X_1(\omega),\cdots,X_n(\omega)\cdots,X_N(\omega)]^T$ and $\mathbf{D}(\omega) = [D_1(\omega),\cdots,D_n(\omega)\cdots,D_N(\omega)]^T$ are vectors of spectrum values at the frequency ω for the signal captured by the n-th microphone, $X_n(\omega)$, and for the ego noise, $D_n(\omega)$, respectively.

$$\mathbf{R}(\omega,\phi) = \mathbf{X}(\omega)\mathbf{X}^*(\omega). \tag{8}$$

$$\mathbf{K}(\omega,\phi) = \mathbf{D}(\omega)\mathbf{D}^*(\omega), \tag{9}$$

where $()^*$ represents the complex conjugate transpose operator and ϕ denotes the orientation of the robot's head. GEVD of $\mathbf{R}(\omega,\phi)$is formulated as follows:

$$\mathbf{K}^{-1}(\omega,\phi)\mathbf{R}(\omega,\phi) = \mathbf{Q}(\omega,\phi)\Lambda\mathbf{Q}^{-1}(\omega,\phi), \tag{10}$$

where Λ is the eigenvalue matrix with $\Lambda_{ii} = \lambda_i$ and $\mathbf{Q}$ is the regular matrix, whose i-th column is the eigenvector $\mathbf{q}_i$. Moreover, we assume that the λ_i and $\mathbf{q}_i$ correspond to the sound sources of interest for $1 \leq i \leq M$ and to the undesired noise sources for $M+1 \leq i \leq N$. $\mathbf{K}^{-1}(\omega,\phi)$ has an effect of whitening the ego noise.

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.

$$\mathbf{P}(\omega,\psi) = \frac{|\mathbf{G}^*(\omega,\psi)\mathbf{G}(\omega,\psi)|}{\sum_{n=M+1}^{N}|\mathbf{G}^*(\omega,\psi)\mathbf{q}_n|}. \tag{11}$$

The peaks occurring in the MUSIC spatial spectrum yield the source locations. The decision on the source locations is made by comparing the sum of the peak powers, $\sum_\omega \mathbf{P}(\omega,\psi)$ to a threshold value T. So far, GEVD-MUSIC was used to detect stationary fan noise only [16]. In our proposed scheme, the predicted templates are used to compute correlation matrices for both fan noise and ego-motion noise on the fly.

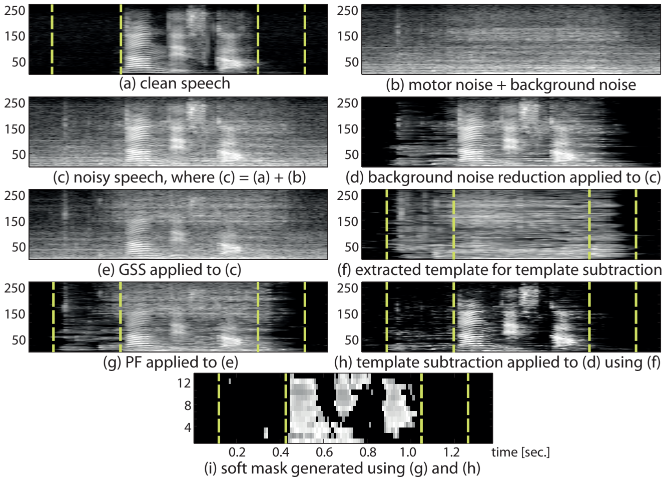
IV. Evaluation

A. ASR System

1) Experimental Settings: We used a circular 8-channel microphone array located on top of the head of a humanoid robot with a height of 1.20 m (See Fig. 1 in [9]). The fan noise is from 180° at a distance of 0.25 m away from the center of microphone array, whereas the 8 arm motors are 0.2-0.5 m and 2 head motors are only 0.1 m away. We recorded (1) random whole-arm pointing behavior as *arm motion* and (2) random head rotation (elevation=[-30° 30°], azimuth=[-90° 90°]) as *head motion*. In terms of noise energy, head motions were 8.4dB higher compared to arm motions. Sensors give the angle of the joints every 5 ms and the length of the audio frames is 10 ms. We used empirical constant values for α=1 and β=0.5 as suggested in [9]. MFM parameters are selected empirically: T=0.25 and σ=10. Except the system depicted in Fig. 2(a), no additional filtering is applied to the incoming data streams.

To generate precise SNR conditions before mixing the noise recording and clean speech, we amplified speech signals based on their segmental SNR. The noise signal consisting of ego noise and environmental background noise is mixed with clean speech utterances used in a typical human-robot interaction dialog and recorded by us. This Japanese word dataset includes 236 words for 4 female and 4 male speakers. The audio data was 8-ch. data convoluted with the transfer functions of the microphone array. Our acoustic model is triphone HMM with 32 mixtures and 2000 states. It was 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 & speaker-open test. We created a matched acoustic model for multi-channel noise reduction (GSS+PF) methods by adding a white noise of $-40dB$. We used 13 static 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. In this experiment, we by-passed SSL to eliminate the mis-localizations with MUSIC due to fan noise and effect of head rotation, and to focus only on the noise suppression performance of our proposed ASR system (unlike in Sec. IV-B.1). Thus, by using transfer functions the position of the speaker is simulated to be 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 0.2sec.

2) Spectrograms and Masks: Fig. 3 gives a general overview about the effect of each processing stage until the masks are generated. In Fig. 3(c), we see a dense mixture of speech (Fig. 3(a)) and motor noise (Fig. 3(b)) with an SNR of -5dB. GSS+PF in Fig. 3(g) reduces only a minor part of the motor noise while sustaining the speech. On the other hand, template subtraction (Fig. 3(h)) reduces the motor noise aggressively while damaging some parts of the speech, where some features of the speech get distorted. The soft mask in (Fig. 3(i)) presents a filter eliminating unreliable and still noisy parts of the speech ($\{T, \sigma\}$=$\{0.5, 5\}$). Furthermore, we observe that features between time intervals of 0.10~0.42 sec. and 1.07~1.27 sec. that are basically composed of only motor noise are given zero weights in the mask except in a few mis-detection cases. The dotted yellow lines in the panels of Fig. 3 indicate the borders of these regions. Note that speech features are located between 0.42~1.07 sec.



(a) clean speech

(b) motor noise + background noise

(c) noisy speech, where (c) = (a) + (b)

(d) background noise reduction applied to (c)

(e) GSS applied to (c)

(f) extracted template for template subtraction

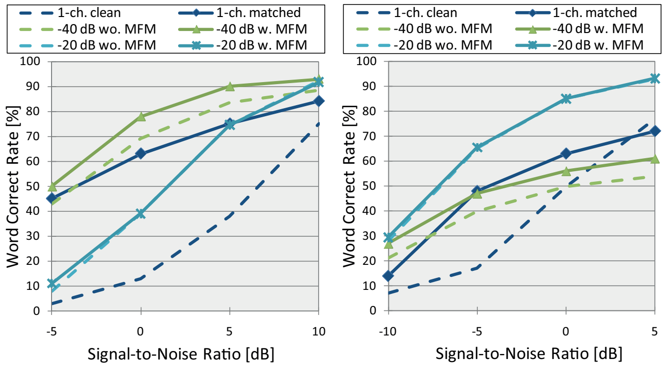
(g) PF applied to (e)

(h) template subtraction applied to (d) using (f)

(i) soft mask generated using (g) and (h)

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

3) ASR Performance: We superimpose white noise of various SNR's and evaluate WCRs with and without MFMs. Fig. 5 shows the ASR accuracies for all methods under consideration. Single-channel results obtained with clean and noise matched acoustic models and without any processing are used as a baseline. In case of arm motion, which is considered as a relatively weaker noise, white noise of the same intensity level used in acoustic model training has shown the best performance. On the other hand, the best ASR accuracy during a head motion with high noise intensity is achieved with an additive white noise of $-20dB$. Based on the results with our robot, where head motion (pan and tilt) noise was louder than background, arm-motion and leg motion noise, we suggest finally that $C1$ and $C2$ in Eq. 1 should be set to $-20dB$ and $-40dB$.



(a) Under arm motion noise

(b) Under head motion noise

Fig. 4. Recognition performance for different types of ego-motion noise

We also observe that the MFT-ASR outperforms the standard ASR without MFMs. Although there is little gain of using MFM for the $-20dB$ white noise (See Fig 5(a) 5(b)), the masks improved the WCR's for all other SNR's during the experiments. While the masks eliminate unreliable speech features contaminated with motor noise, they can also compensate the erroneous effects of voice activity detection due to additive motor noise that contains a large portion of energy. They prevent mis-detection of motor noise as speech, when the speech has not started yet, or is already over.

B. SSL System

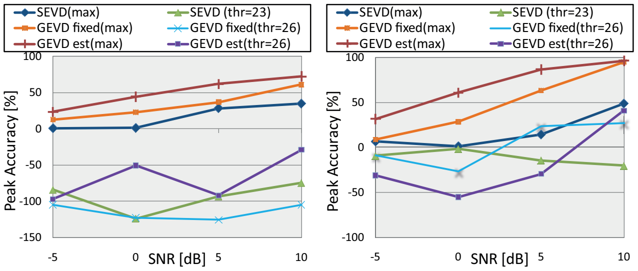
1) Experimental Settings: We compare three SSL techniques: (1) SEVD-MUSIC, (2) GEVD-MUSIC with fixed noise Correlation Matrix (CM) (averaged over 2,000 frames) and (3) proposed method, called GEVD-MUSIC with instantaneously estimated noise CMs. The real-world experiments are conducted for two conditions:

E1) The robot moves its arms randomly (fan noise + arm motion noise)

E2) The robot moves its arms and head randomly (fan noise + arm & head motion noise + head rotation effect)

The resolution of the steering vectors is 5°. The sound source is located 1 meter away at 0° relative to the body of the robot for all experiments. Two types of signals with varying SNR values ranging from -5~10dB are played from a loudspeaker for one minute each: a sinusoidal signal with a fundamental frequency of 600Hz and a white noise signal. Our evaluation criteria are Mean Localization Error (MLE) [°] and the Peak Accuracy (PA [%]) for different threshold values:

$$PeakAccuracy = 100\frac{\#Frames - \#Subst. - \#Del. - \#Ins.}{\#Frames}. \tag{12}$$



(a) PA during arm motion

(b) PA during arm & head motion

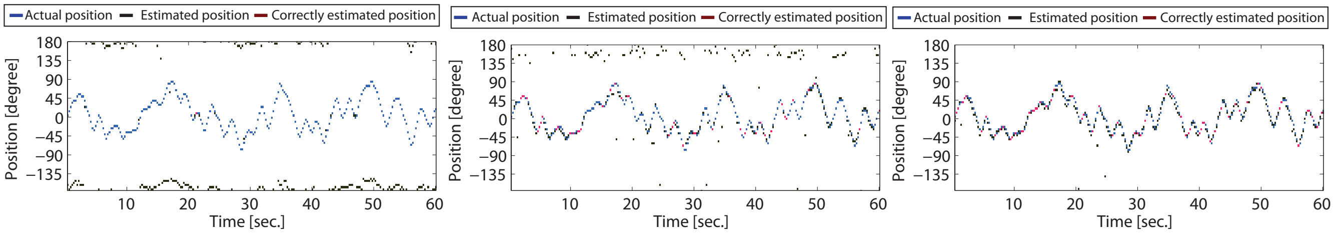
Fig. 6. Peak Accuracy curves for all three methods

2) SSL Results: Tab. I shows that GEVD with estimated noise templates shows superior performance in terms of MLE compared to the other methods in E1 and E2, and almost the same performance like GEVD-fixed in a stationary robot (fan noise only). Generally, SEVD-MUSIC is unable to detect the peak of the desired signal due to the loud fan noise. GEVD-MUSIC with fixed noise CM performs well for fan noise only, and fairly for E1, in which the orientation of the fan noise does not change. The trained CM is still able to suppress the fan noise at a fixed position, however the arm motion noise degrades the performance. In E2, on the other hand, the proposed method is the only method that can eliminate the dynamic noise changes in the spatial spectrum of MUSIC (See Fig. 5).

TABLE I

MEAN LOCALIZATION ERROR (MLE [°]) RESULTS FOR DIFFERENT METHODS

		Fan noise only			*E1) Fan + arm motion noise*			*E2) Fan + arm & head motion noise*		
Signal type	**SNR**	**SEVD**	**GEVD fixed**	**GEVD est**	**SEVD**	**GEVD fixed**	**GEVD est**	**SEVD**	**GEVD fixed**	**GEVD est**
Sinusoidal signal with $f_f = 600Hz$.	*-5*	122.5	69.9	59.1	142.2	93.62	15.8	151.5	82.0	35.2
	0	137.3	**5.1**	**8.4**	164.8	69.79	**9.5**	148.3	56.5	**7.5**
	5	34.3	**5.0**	**1.9**	105.4	**12.0**	**7.8**	31.4	20.1	**7.1**
	10	**5.0**	**5.0**	**0.7**	20.4	**5.9**	**5.9**	26.3	**11.6**	**4.5**
White noise	*-5*	171.5	165.9	161.4	170.5	148.2	162.3	155.1	139.9	116.7
	0	170.6	148.5	151.3	170.6	138.3	95.8	155.2	135.6	101.8
	5	170.3	**1.3**	**1.4**	168.9	35.2	**7.9**	153.6	122.0	32.9
	10	149.2	**0.6**	**0.0**	147.5	**4.1**	**3.9**	146.2	48.1	**12.0**



(a) SEVD

(b) GEVD fixed

(c) GEVD est

Fig. 5. Prediction of positions based on the highest peak of MUSIC spectrum in each frame during random arm and head motion (E3).

We also assess the methods in terms of PA. Fig. 6 illustrates the performance of each method for two different cases: *thr* shows the results obtained with an optimum threshold value, whereas *max* only takes the largest peak into account, thus the deletion and insertion errors in Eq. 12 are automatically omitted. The proposed method outperforms the others in case the maximum peak is selected as the estimated position of the sound source. When a threshold value is used, the performance drops significantly due to the increased insertion errors such as in Fig. 6(a).

3) Discussion: In SSL systems, the number of sound sources (M) and threshold values are the most crucial key points for performance. When the number of sound sources is unknown, a strategy based on a fixed threshold is practical such as in SEVD and GEVD-fixed methods. However, a fixed threshold value for GEVD with estimated noise CM is difficult to determine, because the power of the MUSIC temporal spectrum fluctuates due to the incorrect template predictions, thus its performance is not stable. One way to make the temporal-directional plane of MUSIC smoother is to estimate the CM $\mathbf{K}(\omega,\phi)$ in a longer time window, but it also degrades the noise reduction and SSL performance. Besides, a consequent tracking operation would have improved the final localization accuracies.

In this work, we were mainly interested in our method's capability of suppressing the MUSIC spectrum of noise and dominant noise peaks. We mostly focused on extracting desired sound's peak, therefore we used the strategy of selecting the M largest peaks by assuming that M is given in advance or detected by another process. However, the details about this detection process and the exact correspondence between the sound sources and peaks are still open questions.

V. CONCLUSIONS

In this paper we presented a method for estimating ego noise as a sequence of discrete templates. We inspected the applicability of the approach to different tasks related to robot audition such as robust ASR and SSL. The validity of the ego noise estimation technique was confirmed by quantitative assessments for both of the applications.

In future work, we plan to integrate both SSL and ASR systems and evaluate the combined system in real time and in the real world. Moreover, by extrapolating the identified patterns, we plan to predict missing motion and noise data and add them into the database in an on-line manner.

REFERENCES

[1] G. Ince *et al.*, "A hybrid framework for ego noise cancellation of a robot", in *Proc. of ICRA*, pp. 3623-3628, 2010.

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

[3] S. Yamamoto *et al.*, "Real-time robot audition system that recognizes simultaneous speech in the real world", in *Proc. of IROS*, 2006.

[4] J.-M. Valin *et al.*, "Enhanced Robot Audition Based on Microphone Array Source Separation with Post-Filter", in *Proc. of IROS*, pp. 2123-2128, 2004.

[5] Y. Nishimura *et al.*, "Speech Recognition for a Robot under its Motor Noises by Selective Application of Missing Feature Theory and MLLR", in *Proc. of SAPA*, 2006.

[6] A. Ito *et al.*, "Internal Noise Suppression for Speech Recognition by Small Robots", in *Proc. of Interspeech*, pp.2685-2688, 2005.

[7] J. Even *et al.*, "Semi-blind suppression of internal noise for hands-free robot spoken dialog system", in *Proc.of IROS*, pp.659-663, 2009.

[8] T. Takahashi *et al.*, "Soft Missing-Feature Mask Generation for Simultaneous Speech Recognition System in Robots", in *Proc. of Interspeech*, pp.992-997, 2008.

[9] G. Ince *et al.*, "Ego Noise Suppression of a Robot Using Template Subtraction", in *Proc. of IROS*, pp.199-204, 2009.

[10] G. Ince *et al.*, "Multi-talker speech recognition under ego-motion noise using Missing Feature Theory ", in *Proc. of IROS*, pp.982-987, 2010.

[11] H.D. Kim *et al.*, "Binaural active audition for humanoid robots to localise speech over entire azimuth range", in *Applied Bionics and Biomechanics*, vol. 6, pp.355-367, 2009.

[12] T. Rodemann *et al.*, "Using Binaural and Spectral Cues for Azimuth and Elevation Localization", in *Proc. of IROS*, pp. 2185 - 2190, 2008.

[13] K. Nakadai *et al.*, "Active audition for humanoid", in *Proc. of National Conf. on AAAI*, pp. 832-839, 2000.

[14] I. Cohen and B. Berdugo, "Microphone array post-filtering for non-stationary noise suppression", in *Proc. of ICASSP*, pp.901-904, 2002.

[15] F. Asano *et al.*, "Localization and extraction of brain activity using generalized eigenvalue decomposition", in *Proc. of ICASSP*, pp. 565-568, 2008.

[16] K. Nakamura *et al.*, "Intelligent sound source localization for dynamic environments", in *Proc. of IROS*, pp. 664-669, 2009.