2011 IEEE/RSJ International Conference on
Intelligent Robots and Systems
September 25-30, 2011. San Francisco, CA, USA

Assessment of Single-channel Ego Noise Estimation Methods

Gökhan Ince, Kazuhiro Nakadai, Tobias Rodemann, Jun-ichi Imura, Keisuke Nakamura, and Hirofumi Nakajima

Abstract— **While a robot is moving, ego noise is generated due to the fans and motors of the robot. Furthermore, a robot is not only subject to the ego noise, but also to the ambient noise of the environment, both having different short-term signal characteristics. Because ego-motion noise generated by the motors is non-stationary, and the BackGround Noise (BGN) is stationary, one single noise estimation method is unable to track the changes in both noise spectra rapidly and accurately. Therefore, we propose to use the combination of two different noise estimation methods adequate for each one of co-existing noise types in a unified framework: 1) a *stationary* noise estimation method called Histogram-based Recursive Level Estimation (HRLE) and 2) a *non-stationary* noise estimation method called Template-based Estimation (TE). In this paper, we evaluate the performance of several single-channel based noise estimation techniques in terms of their prediction accuracy and quality of the speech signals enhanced by spectral subtraction methods. The experimental results show that our system, compared to the conventional single-stage noise estimation methods, achieves better performance in attaining signal quality and improving word correct rates.**

I. Introduction

In Automatic Speech Recognition (ASR) systems of mobile robots, the performance degrades drastically in the case of adverse environments with low Signal-to-Noise Ratio (SNR) and the robot's own noise, the "ego noise", which consists of the contributions of various noise sources, such as stationary and diffuse fan/hardware noise and rather non-stationary and directional motor (ego-motion) noise. The signal quality and ASR accuracy can be improved by applying a noise reduction algorithm to the degraded speech.

Conventional adaptive (i.e., Kalman) filtering techniques are mostly not suitable for this noise because they are computationally expensive, require either noise or speech modeling, suffer from adaptation delays and divergence can be harmful. Generally, microphone array-based multi-channel noise reduction methods [1]-[4] demonstrate good performance by attenuating the interfering sound sources, if the sound sources are directional. However, they cannot cope with diffuse type of additive noise, such as static robot/computer/air conditioner noise and partially non-directional ego-motion noise (due to the effects like multi-path propagation, dispersion, dissipation and diffusion inside the covers of the robot). Alternatively and/or complementary to multi-channel approaches, many single-channel speech enhancement algorithms have been developed based on Spectral Subtraction (SS), Wiener Filtering (WF) or Minimum Mean Square Estimation (MMSE) [5]-[7]. All these methods estimate the power spectrum of clean speech using the power spectrum of noisy speech, where the noise estimate plays a major role for the quality of the enhanced signal [8]-[10]. Among these methods, Minima-Controlled Recursive Average (MCRA) [9] is one of the most popular noise estimation methods used in robot audition. It tracks the minimum noise level in the spectrum of the noisy speech and the noise spectrum estimate is obtained from the silent segments of noisy speech using the Voice Activity Detection (VAD) algorithms. The problem is that most VADs are based on SNR levels and make the restricting assumption that the noise is stationary (steady-state). However, a mobile robot is intended to move to various places over time. Therefore, even if the current environment changes, the algorithms should achieve good performance without any manual tuning, such as parameter optimization or learning processes for the new environment. Nakajima *et al.* [10] proposed a noise estimation method, called Histogram-based Recursive Level Estimation (HRLE), which calculates a time-continuous histogram of sound levels in real-time. It shows better performance than MCRA in adapting to the dynamical changes in the environment. Besides, it has the advantage over MCRA that its parameters are tuning-free and does not depend on the SNR-based thresholding.

A robot audition system requires a noise estimation method that can also operate in the presence of non-stationary noise, where the spectral characteristics of the noise changes constantly such as observed in ego-motion noise of a robot. Most of the noise estimation techniques fail in this case, because they are neither able to discriminate non-stationary noise from speech, nor fast enough to track the rapidly changing noise in every frame. Several researchers tackled this problem by using spectral templates recorded in advance [11]-[13]. There are two types of template structures that can be used in Template Estimation (TE): Blockwise template and parameterized template. *Blockwise template* represents the noise spectrum that arises during a complete motion (e.g. trajectory of a single joint,

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}

Hirofumi Nakajima is with Honda Research Institute Japan Co., Ltd. (currently with Kogakuin University) nakajima@cc.kogakuin.ac.jp

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, and Jun-ichi Imura are with Dept. of Mechanical and Environmental Informatics, Tokyo Institute of Technology 2-12-1-W8-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan imura@mei.titech.ac.jp

978-1-61284-455-8/11/$26.00 ©2011 IEEE

trajectory of multiple joints, a motion primitive, a gesture). Using motion commands, the pre-recorded correct noise template similar to the recent motion was selected from a repository of motions and aligned in time according to the current spectrum of the noisy speech. The drawback of this approach is that it cannot cope with dynamic changes of the motion trajectories in time and misalignment is unavoidable [11]. Ito *et al.* [12] proposed a frame-based *parameterized template* prediction technique using an artificial neural network to cope with unstable walking noise of a robot. The trained network had to predict noise spectra for each frame from angular velocities of the joints of the robot. Ince *et al.* [13] extended the idea further by using a template database and Nearest Neighbour (NN) search to extract the estimated templates from the database, because approximate search strategies are more appropriate to estimate the templates from a huge repertoire of robot motions and make an online learning system easier [14].

The strengths of the template-based estimation method are that it is not SNR-dependent, it is not prone to VAD errors and adaptation latency is zero theoretically. However, the template estimation method cannot perform adaptation to the overall (ego and background) noise in an environment with changing noise conditions. It can only reproduce the templates that exist in the database, thus it reflects the ambient noise conditions in the offline training session only. To sum up, there are two major drawbacks: 1) constantly growing database of templates, and 2) incapacity of coping with changing environmental noise in real world. Whereas the former problem can be dealt with an automated incremental learning algorithm proposed in [14] and will not be addressed here, tackling the latter drawback is the primary target of this paper.

In this paper, we comparatively examine the capabilities and performances of several noise estimation methods such as MCRA, HRLE and TE under ego noise of a robot. Since a speech enhancement system solely based on TE cannot cope with the changing environmental noise or an MCRA/HRLE-based noise reduction system is unable to eliminate non-stationary noise, as our contribution, we propose to concatenate the two stages to obtain a unified preprocessing framework for a robot audition system. Our main goals will be (1) to improve the results obtained with objective performance criteria such as Normalized Noise Estimation Error (NNEE), SNR and Log-Spectral Distortion (LSD), and (2) to increase the robustness of other speech processing applications to noise (e.g. ASR).

II. Single-channel Noise Reduction

Suppose an input signal $y(t)$ of time sample t is given such as:

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

where $x(t)$ is a target signal and $n(t)$ is a noise signal. Noise estimation and reduction algorithms operate in the time-frequency (spectrogram) domain. The complex input spectrum $Y(k,l)$ of frequency bin k and time frame l is obtained from

$$Y(k,l) = \sum_{t=0}^{t=W-1} y(t+lM)w(t)\exp\{j(2\pi/W)tk\}, \tag{2}$$

where W is the window length, M is the shift length and $w(t)$ is the window function. The power input spectrum calculated as $|Y(k,l)|^2$ is used to estimate noise power spectrum $\lambda(k,l)$ (defined as $\lambda(k,l) = |\hat{N}(k,l)|^2$) from $|Y(k,l)|^2$ in the noise estimation process. A noise reduction process can be divided into two consequent processes: gain calculation and spectral filter. The gain calculation process calculates the optimum gain $G(k,l)$ that yields the estimated target spectrum as

$$\hat{X}(k,l) = G(k,l)Y(k,l). \tag{3}$$

The equation for computing $G(k,l)$ is derived from the reduction method, e.g., in case of SS [5]:

$$G_{SS}(k,l) = \sqrt{\max\left[\frac{|Y(k,l)|^2 - \lambda(k,l)}{|Y(k,l)|^2}, \beta\right]}, \tag{4}$$

where $\lambda(k,l)$ shows the estimated noise spectrum, *max* shows the maximum value calculation and β is the flooring parameter. Since the noise reduction performance is strongly affected by the quality of $\lambda(k,l)$, the noise estimation method is very important for noise reduction. Fig. 1 shows the general configuration for single-channel noise reduction.

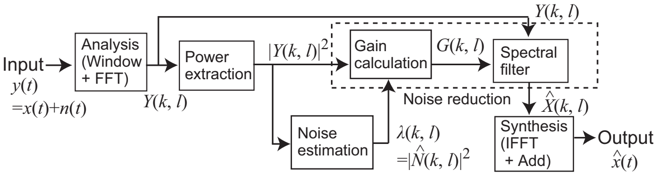

Fig. 1. General configuration for single-channel noise reduction

A. Noise Estimation Methods

1) Minima-Controlled Recursive Average (MCRA): MCRA [9] estimates the noise power as the averaged power of noise periods detected by level-based VAD. First, smoothed input power spectrum $S(k,l)$ is calculated as:

$$S(k,l) = \alpha_s S(k,l-1) + (1-\alpha_s)S_f(k,l), \tag{5}$$

$$S_f(k,l) = \sum_{u=-w}^{w} b(u)|Y(k-u,l)|^2, \tag{6}$$

where α_s, w and $b(u)$ are smoothing parameters. With $S(k,l)$, MCRA calculates the minimum noise spectrum $S_{min}(k,l)$ using the Minimum Tracking method [8], where $S_{min}(k,l)$ is updated for every L frames. The voice activity flag $I(k,l)$, which equals 0 for noise periods and 1 for speech periods, is decided as

$$I(k,l) = \begin{cases} 1 & S_f(k,l) > \eta S_{min}(k,l) \\ 0 & \text{otherwise}, \end{cases} \tag{7}$$

where η is a threshold parameter. The final estimated noise power $\lambda_{MCRA}(k,l)$ is calculated as

$$\begin{aligned}\lambda_{MCRA}(k,l+1) = \lambda_{MCRA}(k,l)p(k,l) + [\alpha_d \lambda_{MCRA}(k,l) \\ +(1-\alpha_d)|Y(k,l)|^2](1-p(k,l)),\end{aligned} \tag{8}$$

where α_d is a smoothing parameter and $p(k,l)$ shows the voice active probability. This $p(k,l)$ is derived as

$$p(k,l) = \alpha_p p(k,l-1) + (1-\alpha_p) I(k,l), \tag{9}$$

where α_p is a smoothing parameter.

MCRA works well when the parameters are optimally adjusted to the acoustic conditions. However, this adjustment is especially difficult in non-stationary environments because the optimum parameters also change accordingly.

2) Histogram-based Recursive Level Estimation (HRLE): HRLE [10] estimates input noise levels as an "x" percentile value L_x value from an input power level histogram. Since HRLE uses recursive averages to obtain temporal histograms, HRLE can adapt smoothly and quickly to the environmental changes. The estimated noise spectrum λ_{HRLE} is obtained as:

$$Y_L(k,l) = 20\log_{10}|Y(k,l)|, \tag{10}$$

$$I_y(k,l) = \lfloor (Y_L(k,l) - L_{min})/L_{step} \rfloor, \tag{11}$$

$$N(k,l,i) = \alpha N(k,l-1,i) + (1-\alpha)\delta(i - I_y(k,l)), \tag{12}$$

$$S(k,l,i) = \sum_{j=0}^{i} N(k,l,j), \tag{13}$$

$$I_x(k,l) = \underset{I}{\operatorname{argmin}} \left[S(k,l,I_{max}) \frac{x}{100} - S(k,l,I) \right], \tag{14}$$

$$\lambda_{HRLE}(k,l) = L_{min} + L_{step} \cdot I_x(k,l). \tag{15}$$

L_{min}, L_{step} and I_{max} are the minimum level, the level width of one bin and the maximum index of the histogram, respectively, x indicates the percentile position, α is the time decay parameter calculated from time constant T_r and sampling frequency F_s as $\alpha = 1 - 1/(T_r F_s)$, $\delta(t)$ shows the Dirac delta function and $\lfloor \cdot \rfloor$ is the flooring function. Especially, x and α influence the estimated level and both are SNR-independent. Furthermore, x value determines how aggressively the noise is estimated. Higher x values are appropriate for non-stationary noises, whereas HRLE with lower x value can capture only stationary noise.

3) Template-based Estimation (TE): TE [13] is a noise estimation method, which is well-suited to capture the dynamic nature of the motion data represented by the sequence of observations. Based on these observations, we are able to associate a discrete time series data (i.e. motion) with another discrete time series data (i.e. ego noise) and predict an arbitrary sequence of associated data. Specifically, this method utilizes encoders attached to the motors of the robot, which measure the angular position of each joint. During the motion of the robot, actual position ($\theta(l)$) information regarding each motor is acquired regularly. Using the difference between consecutive sensor outputs, velocities ($\dot{\theta}(l)$) and accelerations ($\ddot{\theta}(l)$) are calculated. Considering that J joints are active, $3J$ attributes are generated. Each feature is normalized to [0 1] so that all features have the same contribution on the prediction. The resulting feature vector has the form of $F(l) = [\theta_1(l), \dot{\theta}_1(l), \ddot{\theta}_1(l), \theta_2(l), \dot{\theta}_2(l), \ddot{\theta}_2(l), \ldots, \theta_J(l), \dot{\theta}_J(l), \ddot{\theta}_J(l)]$. In the template generation (database creation) phase, one feature vector is thereby assigned to the current noise spectral vector $|Y(l)|^2$ and used to label the instantaneous noise fragment; this data block $T(l) = [F(l) : |Y(l)|^2]$ is called a *parameterized template*. The goal is to create a large noise template database for all desired motions.

During the prediction phase, a nearest neighbor search in the database is conducted for the best matching template of motor noise for the current time instance (frame at that moment) using its feature vector label. The estimated noise power, $\lambda_{TE}(k,l)$, is used to compute the gains as in Eq. (4). Due to the short distance between the motors and microphone array, we assume that the reverberation time of ego noise inside the robot is shorter than one frame ($< 10ms$) and thus represented inside the current template. The motion effects are also modeled in the templates.

B. Proposed Configuration for Ego Noise Reduction

Mobile robots are deployed to environments with (possibly changing) background noise, ego noise and speech. Fig. 2 shows an example of a noise spectrogram that is recorded in this kind of environment. Noise estimation methods based on recursive averaging cannot adapt to the ego-motion noise rapidly (see Fig. 3) because motions of different joint combinations produce different noise spectra. On the other hand, the frame-by-frame based estimation method, TE, is fairly accurate in reconstructing any type of noise that can be reproduced. Ego-motion noise falls into this noise category because the duration and spectral power of the motor noise signals do not change drastically for the same type of motions (Fig. 4). The main drawback of this method is that it cannot perform adaptation to ego-motion noise in an environment with changing noise conditions. It can only reproduce the templates that exist in the database, thus it reflects the noise conditions in the training session only. Considering that

$$N(k,l) = N_s(k,l) + N_n(k,l), \tag{16}$$

where $N_s(k,l)$ and $N_n(k,l)$ denote the stationary and non-stationary portions of the overall noise, we propose to use the stationary and non-stationary noise estimation methods in series as in Fig. 5 for the training of the template database. As a consequence,a unified framework for noise estimation consisting of two parallel and independent processes as in Fig. 6 is created. While recursive averaging takes care of $N_s(k,l)$ the background and stationary portion of ego noise (i.e. fan/hardware noise), TE with a template representation of $T(l) = [F(l) : |\hat{N}_n(l)|^2]$ tackles the remaining non-stationary noise portion of motor noise $N_n(k,l)$.

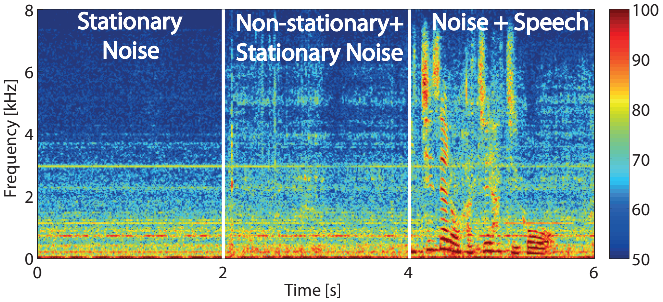

Fig. 2. Noisy spectrogram

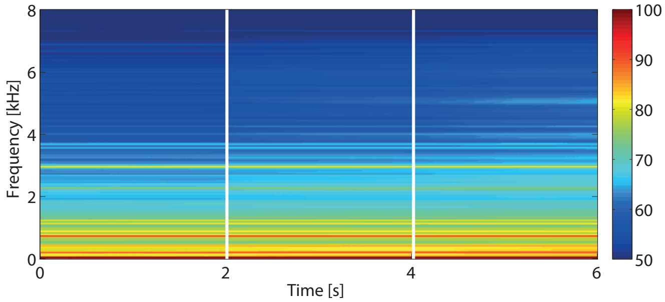

Fig. 3. Estimated noise spectrogram by HRLE with a rather short time constant $T_r = 1sec$.

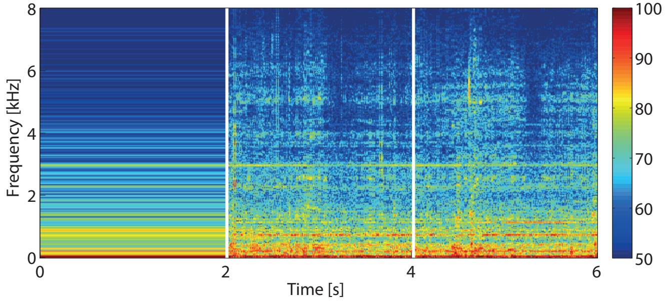

Fig. 4. Estimated noise spectrogram by TE

III. Evaluation

In this section, we first assess the estimation and suppression capabilities of MCRA, HRLE, TE by individually applying them to the noise signals consisting of ego noise and environmental background noise. Then the performance of proposed method with different combinations (i.e., HRLE+TE, MCRA+TE) as explained in Sec. II-B is evaluated. We also test the performance of stationary noise reduction after applying TE-based SS to obtain comparative results. Using a humanoid robot developed by Honda, one set of noise and joint status data (200 seconds long) for training, and three sets of similar data (100 seconds long) for testing are collected during a continuous head motion of 2 Degree of Freedoms (DoF) and arm motion of 4 DoF (see Fig. 7). 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 0.2 sec. The performance of all methods are compared under 4 different SNR conditions for the same signal segments as in Fig. 2. Condition (1)-(2): Noise energy is fixed, speech signals are amplified to yield $SNR_{(1)} = 3dB$ and $SNR_{(2)} = -3dB$; Condition (3)-(4): Gaussian white noise is added to (2) to represent changing conditions of static BGN with $SNR_{(3)} = -3.1dB$ and $SNR_{(4)} = -3.2dB$. The parameters of the HRLE and MCRA are selected appropriately for non-stationary noise estimation [9],[10] and are given in Tab. I. A minor spectral floor $\beta = 0.1$ is used in the SS stage.

TABLE I

Parameter settings for MCRA and HRLE

MCRA	HRLE
$\alpha_d = 0.95$, $\alpha_p = 0.2$	$L_{min} = -200dB$, **wo. TE: / w. TE:**
$\alpha_s = 0.8$, $L = 125$	$L_{step} = 0.2dB$, $x = 50\%/20\%$
$w = 1$, $\lambda_{MCRA} = 5$	$I_{max} = 2000$, $T_r = 1sec/10sec$

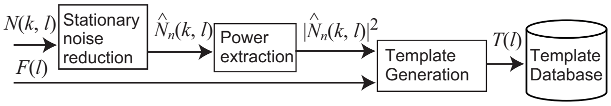

Fig. 5. Template database generation in the offline training session

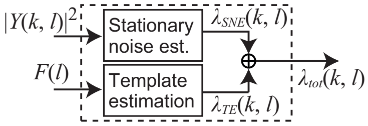

Fig. 6. Unified noise estimation framework

A. Evaluation Criteria

1) 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:

$$\bar{\varepsilon} = \frac{1}{L}\sum_{l=1} 10\cdot\log_{10}\left(\frac{\sum_{k=0}^{M}||N(k,l)|^2 - |\hat{N}(k,l)|^2|)}{\sum_{k=0}^{M}||N(k,l)|^2|}\right), \tag{17}$$

where L is the number of frames.

2) Segmental SNR: The average of the SNR values is calculated for segments of audio data such as:

$$SNR = \frac{1}{L}\sum_{l=1}^{L} 10\cdot\log_{10}\left(\frac{\sum_t x^2(t)}{\sum_t \left(x(t)-\hat{x}(t)\right)^2}\right). \tag{18}$$

3) Log-Spectral Distortion [6]: This evaluation measure computes the reconstruction error of the clean speech by comparing the enhanced speech signal $\hat{X}(k,l)$ with the original speech $X(k,l)$ in the log domain as follows:

$$LSD = \frac{1}{L}\sum_{l=1}^{L}\left(\frac{1}{K}\sum_{k=1}^{K}\left[\mathscr{L}X(k,l) - \mathscr{L}\hat{X}(k,l)\right]^2\right)^{1/2}, \tag{19}$$

where $\mathscr{L}X(k,l) \triangleq \max\{20\log_{10}|X(k,l)|, \delta\}$ is the log spectrum confined to about 50 dB dynamic range, hence $\delta = \max_{k,l}\{20\log_{10}|X(k,l)|\} - 50$.

4) Ideal Estimation: In order to evaluate the accuracy of the noise spectrum estimation, ideal gain $G_i(k,l)$ is computed from the the original noise spectrum in the training session and used in Eq. (3). Note that $G_i(k,l)$ is still subject to negligible errors caused by approximation strategy of the fast NN search[1].

5) 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 Japanese word dataset includes 236 words for 4 female and 4 male speakers. We used a clean acoustic model 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. 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.

[1] http://www.cs.umd.edu/˜mount/ANN/

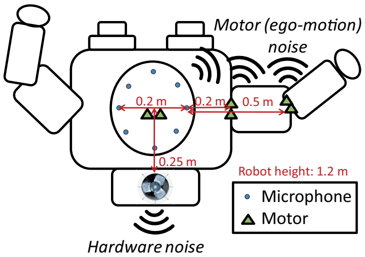

Fig. 7. Experimental setup

B. Results

The *estimation performance* of all methods are given in Tab. II. In all conditions of stationary noise, TE performed worse than other methods because this method is not suitable for estimating the stationary noise. It yields the lowest estimation error for (1) and (2) in the presence of non-stationary noise because the test data was recorded in the same room. However, unfamiliar BGN conditions such as in (3) and (4) degrade the accuracy of the TE because the portion of the stationary noise becomes more dominant in the overall noise energy compared to the portion of the ego noise. Furthermore, HRLE outperforms MCRA especially in the presence of ego noise.

TABLE II

NOISE ESTIMATION PERFORMANCE FOR ALL METHODS

SNR	$\bar{\varepsilon}$ for given segment	HRLE	MCRA	TE
(1) 3dB	Stationary noise	-5.81	**-6.26**	-5.08
	St. + Non-st. noise	-4.61	-4.38	**-4.89**
	Total noise + Speech	-4.92	-4.68	**-5.06**
(2) -3dB	Stationary noise	-5.81	**-6.26**	-5.08
	St. + Non-st. noise	-4.61	-4.38	**-4.89**
	Total noise + Speech	-4.84	-4.63	**-5.06**
(3) -3.1dB	Stationary noise	**-7.96**	-7.12	-4.95
	St. + Non-st. noise	**-6.3**	-5.75	-4.95
	Total noise + Speech	**-6.71**	-5.63	-5.11
(4) -3.2dB	Stationary noise	**-8.87**	-8.03	-4.52
	St. + Non-st. noise	**-7.1**	-6.64	-4.76
	Total noise + Speech	**-7.42**	-6.45	-4.92

We evaluate the *noise reduction performance* by using the system depicted in Fig. 1. This time, we also use the combinations of several estimators in series (labeled as *method A* + *method B*, e.g. HRLE+TE) for database generation and compare their results to the baseline results (i.e., No Processing, *NP*). In this case, we use the the settings in Tab. I indicated with "*w. TE*" because they are more appropriate for stationary noise estimation. As Tab. III demonstrates, TE achieves the smallest LSD and largest WCRs among all methods in the conditions (1) and (2). A substantial improvement of 30.4 points in WCR is achieved especially for $-3dB$. In terms of SNR, only HRLE+TE can outperform TE. In general, HRLE creates less distortion in speech (LSD), thus, achieves higher recognition rates (WCR) compared to MCRA. We also observe that using the stationary noise estimation techniques rather as a secondary step after TE, such as, TE+HRLE or TE+MCRA, does not improve the quality of the refined speech any better than when they are used as a primary step.

Tab. IV provides not only ideal results like, *SNR*, *LSD* or *WCR*, but also normalized noise spectrum errors for stationary noise, st. noise+non-st. noise, total noise + speech, resp. $\bar{\varepsilon}_{P1}$, $\bar{\varepsilon}_{P2}$, $\bar{\varepsilon}_{P3}$. The performance reduction for the combined methods is due to errors in the nonlinear noise reduction operation prior to database generation (see Fig. 5). By making just a small compromise in the accuracies as shown in Tab. II and IV, the framework of HRLE+TE will later provide adaptivity to the system and achieve even better results in changing background noise.

TABLE III

EGO NOISE REDUCTION PERFORMANCE FOR ALL METHODS

Estimation Method	$SNR_{(1)} = 3dB$			$SNR_{(2)} = -3dB$		
	SNR	*LSD*	*WCR*	*SNR*	*LSD*	*WCR*
NP	3.00	9.7	78	-3.0	11.2	28.3
MCRA	3.90	9.49	83.2	-1.38	10.8	44.5
HRLE	3.96	8.94	84.1	-1.2	10.2	47.2
TE	5.49	**8.51**	**87.4**	2.05	**8.73**	**58.7**
MCRA+TE	5.85	8.61	79.7	1.85	8.95	51.6
HRLE+TE	**6.02**	8.66	86.2	**2.74**	8.88	54.8
TE+MCRA	5.28	8.74	84.5	1.44	9.48	53.9
TE+HRLE	5.22	8.91	85.6	1.41	9.3	55.6

TABLE IV

IDEAL NOISE ESTIMATION AND REDUCTION PERFORMANCE FOR $SNR_{(2)} = -3dB$

Method	*SNR*	*LSD*	*WCR*	$\bar{\varepsilon}_{P1}$	$\bar{\varepsilon}_{P2}$	$\bar{\varepsilon}_{P3}$
TE	2.67	7.32	92.6	-12.1	-86.5	-88.0
HRLE+TE	3.84	7.43	92.3	-5.80	-9.22	-8.36
TE+HRLE	2.56	8.19	89.4	-6.06	-12.42	-12.54

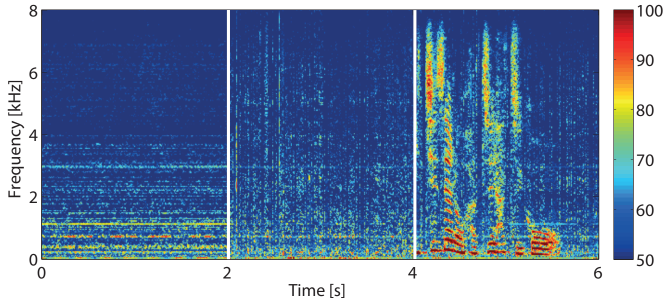

Fig. 8. Refined spectrogram by TE-based SS ($SNR = 3dB$)

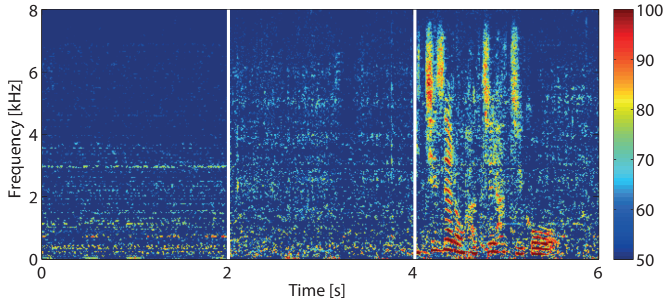

Fig. 9. Refined spectrogram by "HRLE+TE"-based SS ($SNR = 3dB$)

By inspecting the spectrograms of the refined signal, $\hat{X}(k,l)$, in Fig. 8 and Fig. 9, we see clearly that TE has difficulty in estimating the stationary noise segment because it relies only on one single template representing the fixed stationary position of the robot. In addition, it creates sharp vertical valleys and peaks in the spectrum, which are caused

TABLE V

EGO NOISE REDUCTION PERFORMANCE FOR ALL METHODS

Estimation Method	$SNR_{(3)} = -3.1dB$			$SNR_{(4)} = -3.2dB$		
	SNR	*LSD*	*WCR*	*SNR*	*LSD*	*WCR*
NP	-3.1	12.6	25.5	-3.2	14	22.8
MCRA	-1.31	11.2	36.8	-1.23	11.9	34.8
HRLE	-0.75	10.0	42.1	-0.76	9.93	38.2
TE	1.97	9.75	51.7	1.77	11.1	44.5
MCRA+TE	1.93	9.75	40.8	1.89	9.73	17.7
HRLE+TE	**2.63**	**9.03**	**52.0**	**2.77**	**9.24**	**46.5**
TE+MCRA	1.44	10.3	50.3	1.44	11	44.1
TE+HRLE	1.52	9.74	47.1	1.53	10.1	42.2

by the smaller attenuations of the frequencies compared to relatively larger attenuations of their neighboring frequencies due to the incorrect estimations or missing templates in the database. This so-called *musical noise* effect is reduced by the HRLE+TE, because HRLE attenuates the spectrum more smoothly and the characteristics of the residual noise resembles less harmful salt-and-pepper noise.

Tab. V shows the results for the simulation of changing ambient noise. We observe that the higher the contribution of the background noise, the more effective MCRA and HRLE methods are. Under conditions (3) and (4), especially HRLE contributes more to cancelling the overall noise by eliminating the background noise. Hence in its combination with TE, TE deals only with the non-stationary part of the overall noise regarding the ego-motion noise. This kind of configuration improves the robustness of the noise suppression system against changes in the environmental noise conditions. Besides ASR, the high *SNR* and low *LSD* results indicate that the estimates can also be used accurately for other speech applications. Fig. 10 shows the resulting spectrogram when TE-based SS is applied for the noise condition (4). As expected, TE cannot cope with the remaining background noise in all frequency bands. On the other hand, the proposed HRLE+TE based SS method can suppress the noise effectively as shown in Fig. 11. The similarity between Fig. 11 and Fig. 9 justifies the importance of the proposed configuration in a typical audition system of a mobile robot and that it achieves a similar suppression performance even if the environment changes.

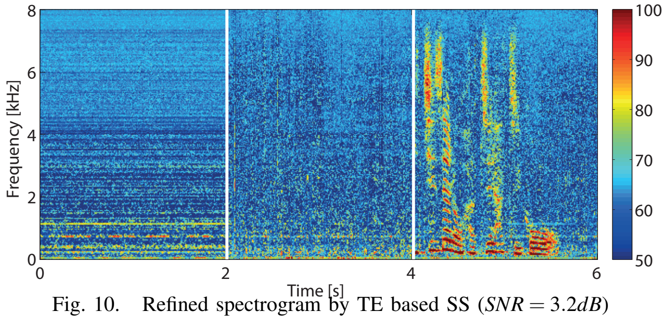

Fig. 10. Refined spectrogram by TE based SS ($SNR = 3.2dB$)

IV. SUMMARY AND OUTLOOK

In this paper we assessed the performance of several single-channel noise reduction methods in the presence of background noise and ego noise, and presented a method exhibiting high robustness even against changing conditions of the environment. The system we proposed combined a stationary noise estimation (HRLE) and non-stationary noise estimation (TE) in a single framework. We showed that our integration method achieves precise estimation of overall noise and a high ASR accuracy under various SNR conditions. Another contribution of this paper was that it provides the underlying basis and configuration of further research advancement in incremenal learning of ego-motion noise templates [14]. In future work, we plan to predict missing motion and noise data by extrapolating the identified patterns, and add them into the database in an online manner.

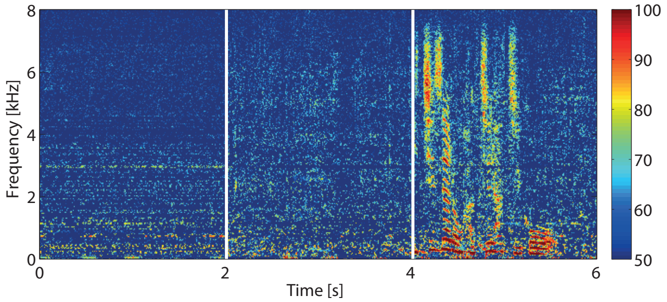

Fig. 11. Refined spectrogram by HRLE+TE based SS ($SNR = 3.2dB$)

REFERENCES

[1] M. Brandstein and D.Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer-Verlag, 2001.

[2] 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.

[3] 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.

[4] 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.

[5] J. Deller, *Discrete-Time Processing of Speech Signals*, IEEE Press, 2000.

[6] J. Benesty, M. M. Sondhi, Y. Huang, *Springer Handbook of Speech Processing*, Springer-Verlag, 2008.

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

[8] R. Martin, "Spectral Subtraction Based on Minimum Statistics", *Proc. Eur. Signal Processing*, 1182-1185, 1994.

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

[10] 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.

[11] 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.

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

[13] 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.

[14] G. Ince, H. Nakajima, K. Nakamura, T. Rodemann, K. Nakadai and J. Imura "Incremental Learning for Ego Noise Estimation of a Robot", to appear in *Proc. of IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, 2011.